

**Die Expression humaner Proteine in der Hefe *Pichia pastoris*:
Hochdurchsatzverfahren und bioinformatische Identifizierung
von Expression-beeinflussenden Sequenzmerkmalen**

vorgelegt von
Diplom-Ingenieur
Mewes Böttner

von der Fakultät für Prozesswissenschaften
der Technischen Universität Berlin
zur Erlangung des akademischen Grades
Doktor der Ingenieurwissenschaften
-Dr. ing.-

genehmigte Dissertation

Promotionsausschuss:

Vorsitzender: Prof. Dr. R. Tressl

Berichter: Prof. Dr. U. Stahl

Berichter: Prof. Dr. D. Mattanovich

Berichter: PD Dr. C. Lang

Tag der wissenschaftlichen Aussprache: 11.03.2004

Berlin 2004

D 83

Danksagung

Herrn Prof. Dr. Ulf Stahl möchte ich für die Möglichkeit danken, diese Arbeit am Institut für Biotechnologie, Fachgebiet Mikrobiologie und Genetik, der Technischen Universität Berlin anzufertigen.

Herzlich bedanken möchte ich mich auch bei Frau PD Dr. Christine Lang für die Betreuung der Doktorarbeit. Gute Anregungen und ein Vertrauen, dass mir viele Freiheiten bei der Durchführung ermöglichte, haben sowohl zum Inhalt der Arbeit als auch zu meinem Spaß bei der Sache beigetragen.

Herr Prof. Dr. Diethard Mattanovich hat sich dankenswerterweise bereiterklärt, diese Arbeit zu begutachten.

Mein ganz besonderer Dank gilt Christina Steffens, die durch sehr gute und geduldige Mitarbeit einen großen Teil zu dem Projekt beitrug, sowie Heiko Krämer, dessen technische Unterstützung ebenfalls gute Beiträge leistete.

Ulrich Hartig danke ich für seine Hilfe bei der PERL Programmierung.

Caterina Holz, Bianka Prinz, Markus Veen und Grit Kasper haben nicht nur die Arbeit kritisch gelesen, sondern mir auch während der Durchführung mit viel Unterstützung zur Seite gestanden.

Alle bis hierhin ungenannten Mitglieder der Arbeitsgruppe, insbesondere Natalia Bolotina, Jeffrey Schultchen, Ralf Rydzewski, Birgit Neukamm und Birgit Baumann trugen mit Keksen, Kommentaren, Methoden und vielen Hilfestellungen zum Gelingen bei.

Die Kollegen in der Proteinstrukturfabrik sorgten für eine gute Atmosphäre, gute Arbeitsmöglichkeiten, Diskussionsbereitschaft und eine schöne Zeit am Heubnerweg.

Ebenfalls danken möchte ich den Mitarbeitern am Fachgebiet Mikrobiologie und Genetik der TU-Berlin für Unterstützung und fruchtbare Diskussionen.

Inhaltsverzeichnis

1 Die Optimierung heterologer Genexpression durch die Veränderung cDNA-spezifischer Parameter 1

1.1	Einflüsse der « codon usage » der kodierenden Sequenz auf die Expression heterologer Proteine.....	2
1.1.1	Der Einfluss der „codon usage“ bei der Expression in <i>E. coli</i>	2
1.1.2	Der Einfluss der „codon usage“ bei der Expression in Hefe.....	5
1.1.3	„Codon usage“ und heterologe Expression in sonstigen Pilzen.....	6
1.1.4	Auswirkungen der „codon usage“ bei Expression in Säugerzellen	7
1.1.5	Strategien zur Vermeidung von Expressionsproblemen aufgrund schlechter „codon adaptation“	8
1.2	Der Einfluss der Nukleotidkomposition auf die Expressionshöhe.....	11
1.2.1	Der Einfluss der Nukleotidkomposition in <i>E. coli</i>	11
1.2.2	Einflüsse der Nukleotidkomposition auf die Expression in Hefe	11
1.2.3	Die Nukleotidkomposition in filamentösen Pilzen	13
1.2.4	Die Auswirkungen AT-reicher Bereiche in Säugerzellen.....	13
1.2.5	Möglichkeiten zur Vermeidung negativer Einflüsse der Nukleotidkomposition	13
1.3	Hemmung der Translation: Einflüsse durch Sekundärstrukturen in der mRNA	15
1.3.1	Die Folgen von Sekundärstrukturen der mRNA in <i>E. coli</i>	15
1.3.2	Der Einfluss von Sekundärstrukturen bei der Expression in Hefe.....	17
1.3.3	Die Vermeidung von Expressionsproblemen aufgrund von Sekundärstrukturen in der mRNA.....	19
1.4	Sonstiges.....	20
1.5	Schlussfolgerungen	20

2 Problemstellung 24

2.1	Entwicklung eines parallelisierten Systems zum „screening“ von Expressionsklonen	25
2.2	Untersuchung der Sequenzen auf Parameter, die mit der beobachteten Expressionshöhe korrelieren	25

3 Material und Methoden 27

3.1	Material	27
3.1.1	Oligonukleotide.....	27
3.1.2	Plasmide	27
3.1.3	Stämme.....	27

3.1.4	Medien.....	28
3.1.5	Puffer, Lösungen, Chemikalien.....	28
3.1.6	Enzyme.....	29
3.1.7	Antikörper	29
3.2	Programme, Server und Datenbanken.....	29
3.3	DNA-Techniken	30
3.3.1	Plasmidisolation aus <i>E. coli</i>	30
3.3.2	Restriktionsverdau.....	30
3.3.3	Reinigung von DNA-Fragmenten	30
3.3.4	Ligation	30
3.3.5	Transformation von <i>E. coli</i>	30
3.3.6	Kolonie-PCR von <i>E. coli</i> Expressionsklonen zur Umklonierung der cDNAs. 31	
3.3.7	Kolonie-PCR von <i>E. coli</i> zur Überprüfung der Klonierung.....	31
3.3.8	Transformation von <i>P. pastoris</i>	31
3.3.9	Kolonie-PCR von <i>P. pastoris</i>	31
3.4	Methoden der RNA-Analyse.....	31
3.4.1	Synthese und Markierung der Sonde	31
3.4.2	Isolation von Gesamt-RNA aus <i>P. pastoris</i>	32
3.4.3	RNA-Gel und Northern-blotting	32
3.5	Methoden der Protein-Analyse.....	33
3.5.1	SDS-PAGE.....	33
3.5.2	Western-Blotting	33
3.5.3	Immunodetektion	33
3.5.4	Metall-Chelat Affinitätschromatographie	33
3.5.5	StrepTactin Affinitätschromatographie.....	34
3.5.6	Kolonie-Blot von <i>P. pastoris</i>	34
3.6	Bioinformatische Methoden.....	34
3.6.1	Erstellung phylogenetischer Bäume.....	34
3.6.2	Ermittlung und Charakterisierung von Sequenzmotiven	34
3.6.3	Quantifizierung AT-reicher Regionen	35
3.6.4	Gesamt GC-Gehalt und GC-Gehalt an dritten synonymen Positionen (GC3s)35	
3.6.5	Messung der „codon usage“	35
3.6.6	Die Verteilung seltener Codone	35
3.6.7	Generelle Proteinkennzahlen.....	35

3.6.8	Proteindegradationssignale.....	36
3.6.9	Bestimmung der Ähnlichkeiten der untersuchten Sequenzen zu annotierten Hefeproteinen mittels BLAST.....	36
3.6.10	Zuordnung der Proteine zu strukturellen „protein superfamilies“, die auch in <i>S. cerevisiae</i> vertreten sind.....	36
3.6.11	Vorhersage von Sekundärstrukturmerkmalen.....	37
3.7	Statistische Evaluierung.....	37
4	Ergebnisse	38
4.1	Optimierung der Expression im Schüttelkolben durch Induktions- und Medienvariation.....	38
4.1.1	Konstruktion und Evaluierung des Expressionsvektors für GFP.....	38
4.1.2	Variation von Zufütterung und Medium.....	39
4.2	Entwicklung eines parallelisierten Systems zum „screening“ von Expressionsklonen.....	42
4.2.1	Korrelation der Proteinexpression mit dem Nachweis der Expressionskassette mittels PCR.....	42
4.2.2	Konstruktion des PSF-Expressionsvektors	43
4.2.3	Klonierung und Expression verschiedener humaner cDNAs.....	44
4.2.4	Evaluierung der Affinitätstags in <i>P. pastoris</i> – Reinigung von Beispielproteinen	50
4.3	Warum werden verschiedene cDNAs derart unterschiedlich exprimiert – die Suche nach Parametern, die die Expression beeinflussen.....	51
4.3.1	Untersuchungen zur transkriptionellen Regulation und mRNA Stabilität.....	51
4.3.2	Ungerichtete Suche nach Sequenzmerkmalen, die mit der Expressionshöhe korrelieren.....	57
4.3.3	Gerichtete Analyse sequenzbasierter Parameter und deren Verteilung auf die Kategorien der Expressionshöhe	61
4.3.4	Zweidimensionale Auswertungen – Kombination verschiedener Parameter ..	71
5	Diskussion.....	73
5.1	Entwicklung eines parallelisierten Systems zum „screening“ von Expressionsklonen und Erstellung einer Sammlung von Expressionsklonen.....	73
5.1.1	Variation von Zufütterung und Medium	73
5.1.2	Zusammenhang zwischen Integration der Expressionskassette und Proteinexpression.....	76
5.1.3	Entwicklung der Expressionskontrolle im 2 mL Maßstab.....	76

5.2	Ursachen für die beobachteten Unterschiede in der Expressionshöhe.....	77
5.2.1	Menge und Stabilität der Transkripte.....	77
5.2.2	Die Qualität der Sequenzinformation.....	79
5.2.3	Ungerichtete Suche nach Sequenzmerkmalen, die mit der Expressionshöhe zusammenhängen.....	80
5.2.4	Die Verteilung sequenzbezogener Parameter	82
5.3	Fazit und Ausblick	92
6	Zusammenfassung	94
7	Literatur	95
8	Anhang.....	111

Abkürzungen

A	Adenin
Abb.	Abbildung
Acc. No.	„accession number“
AmpR	Ampecillin-Resistenz Gen
AOX1	Alkoholoxidase 1
APS	Ammoniumpersulfat
AS	Aminosäure
BLAST	„basic local alignment search tool“
C	Cytosin
CAI	„codon adaptation index“
Cole	Replikationsursprung für <i>E. coli</i>
C-Quelle	Kohlenstoffquelle
Da	Dalton
d.h.	das heißt
DIG	Digoxigenin
DNA	Desoxyribonukleinsäure
DTT	Dithiothreitol
dNTP	Desoxyribonukleotid
ER	Endoplasmatisches Retikulum
E-value	„expect value“
g	Gramm
G	Guanin
GC3s	GC-Gehalt an dritten synonymen Positionen
GFP	"green fluorescent protein"
h	Stunde
HMM	„hidden Markov model“
HRP	„horseradish peroxidase“
IMAC	„immobilized metal affinity chromatography“
l	Liter
LB	Luria broth
m	Meter
M	molar
MEME	„multiple Em for motif elicitation“
min	Minute
MOPS	Morpholinopropan-sulfonsäure
mRNA	"messenger"-RNA
n.a.	nicht analysiert
Nc	„number of effective codons“
nt	Nukleotid
OD	Optische Dichte
ORF	„open reading frame“

ori	"origin of replication"
PBS	"phosphate buffered saline"
PCR	„polymerase chain reaction“
Pfam	„Protein families database of alignments and HMMs“
pI	isoelektrischer Punkt
PMSF	Phenylmethylsulfonyl-fluorid
PSF	Protein Struktur Fabrik
P-value	„probability value“
RNA	Ribonukleinsäure
rRNA	ribosomale RNA
RT	Raumtemperatur
s	Sekunde
SCOP	„structural classification of proteins“
SDS	Natriumdodecylsulfat
SGD	Saccharomyces Genome Database
SSC	"saline buffered sodium citrate"
SMART	„Simple Modular Architecture Research Tool“
T	Thymin
TAE	Tris-Acetat-EDTA
TEMED	N, N, N', N'-Tetra-methylethylethylendiamin
Tris	Tris (hydroxymethyl) aminomethan
tRNA	"transfer"-RNA
U	Uracil
UTR	untranslatierte Region
UV	ultraviolette Strahlung
V	Volt
x g	Erdbeschleunigung
YE	"yeast extract"
YNB	"yeast nitrogen base"
z.B.	zum Beispiel

1 Die Optimierung heterologer Genexpression durch die Veränderung cDNA-spezifischer Parameter

Der heterologen Expression von Proteinen kommt wachsende Bedeutung zu. Sowohl in biochemischen Studien als auch in biotechnologischen Produktionsprozessen ist diese Technologie häufig Voraussetzung für die Gewinnung größerer Mengen reinen Proteins. Vor allem im industriellen Maßstab wird, bedingt durch die schnell wachsende Anzahl von entsprechenden Molekülen in den Entwicklungsabteilungen der Pharmafirmen, weltweit ein erheblicher Mangel an Produktionskapazitäten prognostiziert (Garber 2001). Ein weit bedeutenderer Flaschenhals dieser Technologie ist, dass sich einige Proteine gar nicht oder nur mit geringen Ausbeuten heterolog exprimieren lassen (z.B. Pikaart und Felsenfeld 1996; Milek et al. 2000; Boettner et al. 2002). Hier besteht Optimierungsbedarf, um zum einen die vorhandenen Produktionskapazitäten effektiver zu nutzen, zum anderen bestimmte Proteine, die sich bis dato nur in zu geringen Mengen herstellen lassen, überhaupt erst in nutzbaren Mengen zur Verfügung zu stellen. Die Möglichkeiten der Prozessoptimierung sind entsprechend der Komplexität der Aufgabe vielfältig. Neben den rein verfahrenstechnischen Optimierungen rund um die Kultivierung des Wirtsorganismus sowie der anschließenden „downstream“ Prozessierung wird hierbei ein steigendes Augenmerk auf die Physiologie des Expressionswirts sowie die genetischen Parameter der rekombinanten Proteinexpression gelegt (z.B. Balbas 2001; Peeters et al. 2001; Punt et al. 2002). Dies kann zum einen zur Expression von Proteinen führen, die vorher gar nicht herstellbar waren, zum anderen aber auch den Anteil des rekombinanten Produktes am Gesamtprotein erhöhen, was nicht nur die Produktausbeute erhöht, sondern auch die sich anschließende Isolation und Reinigung vereinfacht.

Die relevanten genetischen Parameter sind vor allem die Art des Expressionsvektors (Kopienzahl, integrative oder episomale Replikation, Art der Selektion etc.), transkriptionell regulatorische Sequenzen (z.B. Promoter und Terminator), Stabilität und Struktur der mRNA, translationell regulatorische Bereiche (Ribosomenbindungsstellen, Translationsinitiationsstellen) bis hin zu Eigenschaften der Sequenz selbst, wie z.B. der „codon usage“.

Für Proteine, deren Herstellung auf genetischer oder wirtsphysiologischer Ebene so stark limitiert ist, dass die Proteinmenge für ein weiteres „downstream processing“ nicht ausreicht, ist das Optimierungspotential von verfahrenstechnischen Ansätzen gering. Neben der Variation des Expressionssystems – Wirtsorganismus sowie Vektorsystem – lassen sich hier

aber Erfolge über die Modifikation der kodierenden Sequenz und/oder der mRNA Sequenz und Struktur erzielen (z.B. Deng 1997; Batard et al. 2000; Sinclair und Choy 2002).

In der folgenden Arbeit wird der Stand des Wissens mit Fokus auf die Optimierung von kodierenden Sequenzen zur heterologen Expression diskutiert.

1.1 Einflüsse der « codon usage » der kodierenden Sequenz auf die Expression heterologer Proteine

Es ist bereits seit längerem und für verschiedene Organismen bekannt, dass synonyme Codone unterschiedlich häufig genutzt werden (Bennetzen und Hall 1982; Sharp et al. 1986). Statistische Analysen zeigen eine starke Korrelation zwischen der Häufigkeit bestimmter Codone und der Expressionshöhe des entsprechenden Gens (Gouy und Gautier 1982; Sharp et al. 1986; Jansen et al. 2003). Das heißt, dass hochexprimierte Gene eine starke Tendenz zur Nutzung nur bestimmter synonymer Codone zeigen. Sowohl für *Escherichia coli* als auch für die Hefe *Saccharomyces cerevisiae* gilt, dass diejenigen Codone überrepräsentiert sind, die von häufiger vorkommenden tRNAs erkannt werden. Dies impliziert einen Vorteil bei der Translation für Gene, die bevorzugte Codone verwenden (Ikemura 1982; Bulmer 1987).

Die Trends in der „codon usage“ sind bei verschiedenen Organismen nicht identisch (Grantham et al. 1980; Ikemura 1985). Daher werden bei einer heterologen Expression Gene, deren „codon usage“ an den Ursprungsorganismus angepasst ist, in einen Expressionswirt eingebracht, der unter Umständen andere bevorzugte Codone hat. Der Grad der Anpassung der „codon usage“ einer gegebenen Sequenz an die des Wirtsorganismus wird „codon adaptation“ genannt. Eine schlechte „codon adaptation“ kann Einfluss sowohl auf die Expressionshöhe als auch auf die Authentizität des exprimierten Proteins haben.

Für verschiedene gebräuchliche Wirtssysteme gibt es hierzu bereits detaillierte Untersuchungen.

1.1.1 Der Einfluss der „codon usage“ bei der Expression in *E. coli*

Die folgenden Beobachtungen wurden zum Einfluss der „codon usage“ auf die Expression in *E. coli* gemacht:

Im Gen für GATA-1 aus Huhn befinden sich im C-terminalen Bereich der kodierenden Sequenz zwei Anhäufungen von jeweils drei bzw. vier aufeinanderfolgenden seltenen Glycin-Codonen (GGG und GGA). Erst eine Mutagenese dieser Codone führt zur Expression des Proteins (Pikaart und Felsenfeld 1996). Im c-Fos Gen aus Maus befinden sich zwei Anhäufungen von seltenen Arginin Codonen, die ebenfalls die Expression in *E. coli*

verhindern. Erst eine gleichzeitige Mutagenese beider Bereiche führt zur Expression. Daraus wurde gefolgert, dass es keine, stetig mit der Zahl der seltenen Codone steigende Hemmung gibt, sondern einen Grenzwert, dessen Überschreitung eine Expression verhindert (Deng 1997).

Mehrere einzeln stehende seltene Codone können ebenfalls einen negativen Einfluss auf die Expressionshöhe haben (Zdanovsky und Zdanovskaia 2000; Acosta-Rivero et al. 2002; Ma et al. 2003). So konnte die Expressionshöhe zweier Gene, die reich an den seltenen Arginin-Codonen AGA und AGG sind, mittels Ko-Überexpression des *E. coli* Gens *argU* – kodierend für die seltene tRNA_{AGA/AGG} – um den Faktor drei bzw. fünf erhöht werden (Acosta-Rivero et al. 2002).

Verschiedene Arbeiten untersuchten den Einfluss der „codon-usage“ im 5'-Bereich der cDNA. So wurde der Einfluss des Codons unmittelbar nach dem Start-Codon mithilfe des lacZ-Proteins als Reporter untersucht (Looman et al. 1987; Stenstrom et al. 2001). Die Insertion aller 64 möglichen Codone an der +2 Position resultiert in einer Variation der Expressionshöhe um den Faktor 15 unter Verwendung einer starken Ribosomenbindungstelle (Looman et al. 1987) und um den Faktor 20 in Kombination mit einer schwachen Ribosomenbindungstelle (Stenstrom et al. 2001). In beiden Untersuchungen gibt es keine Korrelation zwischen der Expressionshöhe des Reporters und der „codon adaptation“ an dieser Position. Der Einfluss bestimmter Codone auf die Expressionshöhe unterscheidet sich zum Teil stark zwischen den beiden Arbeiten. Zumindest im zweiten Fall bleibt die Halbwertszeit der Transkripte unbeeinflusst. Als Gründe für die Widersprüche zu Looman und Mitarbeitern (1987) werden mögliche Wechselwirkungen mit der unterschiedlichen Shine-Dalgarno Sequenz vorgeschlagen (Stenstrom et al. 2001).

Die Expressionshöhe von humaner Glutathiontransferase konnte durch ein Screening von Klonen, die durch ungerichtete stille Mutationen in den ersten zehn Codonen der 5'-Region erzeugt wurden, erhöht werden. Interessanterweise führte diese Mutagenese nicht zu einer im Hinblick auf die „codon usage“ optimierten Sequenz. Von den zehn 5'-Codonen in der höchstexprimierenden Mutante gehörten vier zu den in *E. coli* nicht bevorzugten (Nilsson und Mannervik 2001). Die Autoren folgern, dass es keine einfache Regel für die Wahl der günstigsten Codone in dieser Region gibt.

Eine weitere Beobachtung bezüglich der Rolle der „codon usage“ im 5'-Bereich wurde anhand verschiedener Versionen von Somatotropin aus Schwein gemacht: bei Proteinen, die aufgrund vektorbasierter Faktoren – starker Promotor, gute Shine-Dalgarno Sequenz – hoch

exprimiert wurden, spielte die „codon usage“ im 5'-Bereich der kodierenden Sequenz keine Rolle (Wang et al. 1993).

In einem Fall konnte ein negativer Einfluss eines seltenen Codons in der +2 Position gezeigt werden (Ma et al. 2003). Es ist allerdings nicht festzustellen, ob dies spezifisch für das 5'-Ende der Sequenz oder ein allgemeiner Einfluss der „codon usage“ ist.

Die Untersuchungen belegen einen Einfluss der „codon usage“ auf die Expression von heterologen Genen in *E. coli*. Dieser Einfluss ist nicht proportional zur Anzahl seltener Codone. Weiterhin lässt sich kein einfacher Zusammenhang zwischen seltenen Codonen im 5'-Bereich der kodierenden Sequenz und der Expressionshöhe belegen. Da die Nukleotidsequenz in diesem Bereich die Expressionshöhe jedoch stark beeinflussen kann, müssen hier andere sequenzbasierte Faktoren eine Rolle spielen.

Ein anderer interessanter und unerwarteter Effekt zeigte sich bei der Untersuchung des Einflusses von seltenen Codonen am 3'-Ende der kodierenden Sequenz von Chloramphenicoltransferase (CAT) in *E. coli* (Gursky und Beabealashvili 1994). Die Einführung von zwei seltenen Arginin-Codonen (AGG) unmittelbar vor dem Stop-Codon führte zu einer bis zu 10-fachen Erhöhung der Proteinmenge. Hier zeigte sich, dass die Menge des CAT-spezifischen Transkriptes gegenüber dem Wildtypgen stark erhöht war. Eine Koexpression der entsprechenden Anticodon-tRNA führte konzentrationsabhängig sowohl zur Abnahme des Transkriptes als auch der Proteinmenge auf Wildtypniveau. Die Autoren postulieren eine Verzögerung der Translation an der Position der seltenen Codone – bedingt durch ein Warten auf die entsprechende tRNA – als Ursache für eine Art Ribosomenstau auf der mRNA, was selbige unzugänglicher für Nukleasen und damit stabiler macht.

Neben einer Beeinflussung der Expressionshöhe wurde eine weitere Auswirkung seltener Codone bei heterologer Expression in *E. coli* beobachtet: so wurde für das seltene Arginin-Codon AGA ein teilweiser Falscheinbau von Lysin beobachtet (Seetharam et al. 1988), der bis zu 42 % des rekombinanten Materials betreffen kann (Calderone et al. 1996). Dieser Effekt wird nicht nur durch gehäuft auftretende, sondern auch durch einzelne seltene Codone verursacht (Calderone et al. 1996). Zwei aufeinanderfolgende seltene Arginin-Codone (AGG) führten zu 50 % fehlerhaft translatiertem Protein (Spanjaard und van Duin 1988). Mehr als eine Leserasterverschiebung ereignete sich bei der Expression von Placenta-Lactogen aus Rind (Kane et al. 1992). Hier erfolgten Leserasterverschiebungen an neun Arginin-Codonen (AGA und AGG). Die Expression der N-terminalen Domäne der p27-Protease aus Herpes-Simplex-Virus führte hauptsächlich zu Protein mit zu großer Masse (78 % des rekombinanten

Proteins) (McNulty et al. 2003). Die ursächlichen Leserastersprünge finden in einer Anhäufung der seltenen Arginin-Codone CGG statt.

Ein entsprechendes Phänomen wurde auch für ein seltenes Codon für Prolin gefunden: hier führte das Codon CCC in der Sequenz für das humane Transferrin ebenfalls zu einer Leserasterverschiebung (de Smit et al. 1994).

Arginin scheint allerdings die problematischste Aminosäure bei der Expression in *E. coli* zu sein, möglicherweise da vier der sechs möglichen Codone in *E. coli* mit einer Frequenz von unter 1 % genutzt werden und somit als „selten“ anzusehen sind (McNulty et al. 2003).

Eine andere Form von „Fehltranslation“ in *E. coli* ist für das Stopcodon TGA beschrieben worden (Lu et al. 1995). Die Expression eines humanen Wachstumsfaktors (PDGF-B) führte neben der korrekten Version hauptsächlich zu einem größeren Protein. Dieses wurde durch Einbau von Tryptophan an der Stelle des Stopcodons TGA und Translation bis zum nächsten, zufälligen TAG-Terminationscodon verursacht. Eine Mutagenese von TGA zu TAG führte zur ausschließlichen Synthese des korrekt terminierten Proteins.

1.1.2 Der Einfluss der „codon usage“ bei der Expression in Hefe

Auch für die Expression in Hefe gibt es Beispiele, bei denen eine ungünstige „codon adaptation“ einen negativen Einfluss auf die Expressionshöhe hat (*S. cerevisiae* und *P. pastoris*: Brocca et al. 1998). Im Unterschied zu einigen der unter 1.1.1 diskutierten Arbeiten, ist ein Einfluss von einzelnen, seltenen Codonen bisher nicht untersucht worden.

Bei der Analyse von 20 *S. cerevisiae* Genen mit verschiedenen mRNA-Halbwertszeiten wurde eine signifikante Häufung von seltenen Codonen in den weniger stabilen Transkripten gefunden (Herrick et al. 1990). Die Einführung von 13 selten benutzten Codonen in das 5'-Ende des Gens für die hefeeigene Pyruvatkinase führte nicht zu einer Verminderung der Translationsrate (Bettany et al. 1989).

Detailliert betrachtet wurde die Rolle der „codon usage“ im 5'-Bereich der kodierenden Sequenz bei heterologer Expression in *S. cerevisiae*:

Unter Nutzung des bakteriellen lacZ-Proteins als Reporter wurde der Einfluss des Codons direkt nach dem Start-Codon untersucht (Looman et al. 1991). 32 verschiedene Codone wurden an dieser Position inseriert, was bis zu einer 5,3-fachen Variation der Expressionshöhe führte. Es war jedoch kein Zusammenhang zwischen der Expressionshöhe und „codon adaptation“ zu erkennen. Die Fusion einer Sequenz, reich an bevorzugten Codonen, an das 5'-Ende der Sequenz kodierend für den vesikulären Monamin-Transporters

aus Ratte führte zu einer Erhöhung der Expression in *S. cerevisiae* (Yelin und Schuldiner 2001).

Zusammengenommen war ein Einfluss seltener Codone am 5'-Ende bisher nicht nachzuweisen. Der positive Einfluss der 5'-Fusion bei Yelin und Schuldiner (2001) resultiert aus einer weit stärkeren Veränderung der Sequenz als es durch die Variation des +2 Codons bei Looman und Mitarbeitern (1991) erfolgt. Es wurden 23 Codone fusioniert, von denen 17 zu den häufig benutzten gehörten (Yelin und Schuldiner 2001).

Leserasterverschiebungen aufgrund von seltenen Codonen sind für heterologe Expression in Hefe nicht wie für *E. coli* erkannt bzw. beschrieben. Grundsätzlich sind Leserasterverschiebungen auch in *S. cerevisiae* ein bekanntes Phänomen. Leserasterverschiebungen als Folge von im +1 Leserahmen hybridisierenden tRNAs – analog zu den oben für *E. coli* dargelegten Phänomenen – wurden für das Retrotransposon Ty beobachtet (zusammengefasst von Farabaugh 1996). Die Leserasterverschiebung ereignet sich an der Sequenz CUU AGG C (für Ty1) oder GCG AGU C (für Ty3) (Pande et al. 1995). Zumindest das Arginincodon AGG im ersten Fall ist ein seltenes Codon in *S. cerevisiae* und Überexpression der entsprechenden tRNA reduzierte die Leserasterverschiebung um den Faktor 43. Dies impliziert einen Mechanismus, der, ähnlich wie in *E. coli*, durch Limitierungen während der Translation bedingt ist (Farabaugh 1996). Die Auswirkung obengenannter Sequenzen auf die heterologe Expression eines lacZ-Reporterkonstruktes untersuchten Pande und Mitarbeiter (1995). Sieben verschiedene Codone an Stelle des Codons AGU in der Sequenz GCG AGU C verursachten Leserasterverschiebungen in bis zu 31 % des Produktes. Allerdings war dies nicht in allen Fällen korrelierbar zur Häufigkeit der Anticodon-tRNA für das eingesetzte Codon.

Weitere Beispiele für mangelnden Expressionserfolg in Hefe sind bis jetzt nicht bekannt. Möglicherweise sind derartige Phänomene in Hefe nicht in erster Linie an seltene Codone geknüpft, sondern benötigen spezifische Erkennungssequenzen wie z.B. Ty-Sequenzen.

1.1.3 „Codon usage“ und heterologe Expression in sonstigen Pilzen

Für andere Pilze als Expressionssystem ist der Einfluss der „codon usage“ bisher nur einmal beschrieben. In dem filamentösen Pilz *Trichoderma reesei* konnte eine Erhöhung der Expression von Xylanase B aus *Dictyoglomus thermophilum* durch Austausch von 20 Codonen mit in *T. reesei* bevorzugten Codone erreicht werden (Te'o et al. 2000).

1.1.4 Auswirkungen der „codon usage“ bei Expression in Säugerzellen

Der Effekt der „codon usage“ auf die heterologe Expression in Säugerzellen wurde in mehreren Arbeiten beschrieben. So konnte die Expression zweier Gene aus *Caenorhabditis elegans* in Säugerzellen um den Faktor 6 – 9 durch Anpassung der „codon usage“ erhöht werden (Slimko und Lester 2003). Ebenfalls positiv auf die Expressionshöhe wirkte sich die Anpassung der vorderen Hälfte der kodierenden Sequenz für einen G-Protein gekoppelten Rezeptor aus *Schistosoma mansoni* aus. In diesem Bereich des nativen Gens befinden sich mehrere Anhäufungen seltener Codone (Hamdan et al. 2002). Eine Steigerung der Expression in COS-Zellen um den Faktor 30 konnte durch „codon adaptation“ eines Gens für ein Allergen aus der Hausstaubmilbe erreicht werden (Massaer et al. 2001). Die Expression von vier HIV Genen in humanen Zellen konnte ebenfalls durch Anpassung der „codon usage“ erhöht werden (Corbet et al. 2000).

Die Änderung von 33 von 83 Codonen für ein Protein des humanen Papillomavirus resultierte in einer neunfachen Erhöhung der Expression in COS-Zellen (Disbrow et al. 2003). Unabhängig davon synthetisierten Cid-Arregui und Mitarbeiter (2003) das Gen für das gleiche Protein aus ausschließlich bevorzugten humanen Codonen. Dies führte zu einer 20 – 100fachen Erhöhung der Expression in humanen Zellen.

Eine „codon adaptation“ eines HIV-Antigens führte zu einer Erhöhung der Expression in humanen, Maus- und Affenzellen. In diesem Fall wird durch die „codon adaptation“ gleichzeitig eine Erhöhung des GC-Gehaltes herbeigeführt. Der Einfluss dieser beiden Parameter wurde jedoch nicht getrennt untersucht (Deml et al. 2001). An diesem Beispiel wurde in einer Folgearbeit der Einfluss von einzelnen, seltenen Codonen analysiert. Das führte zu dem Schluss, dass hier das Verhältnis von seltenen zu bevorzugten Codonen entscheidend ist (Kofman et al. 2003).

In den meisten Arbeiten wurden Änderungen der resultierenden Sequenzen bezüglich des GC-Gehalts, der mRNA Sekundärstrukturen oder anderen Parametern nicht diskutiert. Zumindest in zwei Arbeiten – interessanterweise zum gleichen Protein – wurden Änderungen bezüglich der mRNA bestimmt: eine Neusynthese des Gens aus Papillomavirus mit ausschließlich bevorzugten Codonen führt zu einer Erhöhung der mRNA Stabilität (Cid-Arregui et al. 2003), eine moderatere Anpassung der „codon usage“ dieser Sequenz führt zu keiner Änderung der Transkript-Menge (Disbrow et al. 2003). Zu beachten ist hier, dass zwei verschiedene Parameter gemessen wurden.

Eine detailliertere Analyse der zusammenspielenden Faktoren wurde am Beispiel des humanen Erythropoietin (EPO) durchgeführt (Kim et al. 1997). Zunächst führte eine

Optimierung der Sequenz in Richtung humaner „codon usage“ zu einer erhöhten Expression. Humane Codone sind tendenziell GC-reich, was im allgemeinen zu einer Erhöhung des GC-Gehaltes der Sequenz als Folge der Codon-Optimierung führt. Vergleichende Analysen mit Varianten des 5'-Endes der kodierenden Sequenz in Kombination mit verschiedenen Promotoren und Signalsequenzen führten zu der Erkenntnis, dass ein „codon usage“ optimiertes Gen mit einem 5'-Bereich, der einen niedrigen GC-Gehalt aufweist, am höchsten exprimiert wird. Die Autoren folgern eine inhibierende Wirkung von Sekundärstrukturen im 5'-Bereich des Gens – selbige sind aufgrund des niedrigen GC-Gehaltes instabiler – die über die seltener Codone dominiert.

1.1.5 Strategien zur Vermeidung von Expressionsproblemen aufgrund schlechter „codon adaptation“

Als hauptsächliche Maßnahme, um die „codon adaptation“ zu verbessern, werden stille Mutationen eingesetzt. In *E. coli* können so Limitierungen, die zu verminderter Expressionshöhe führen, überwunden werden (Pikaart und Felsenfeld 1996; Deng 1997; Ma et al. 2003). Hierdurch konnte die Expression in Einzelfällen von „nicht detektierbar“ auf bis zu 20 % des Gesamtproteins erhöht werden (Deng 1997).

Auch für Hefe sind Fälle beschrieben, in denen eine Synthese der kodierenden Sequenz mit stillen Mutationen zur Expression des Proteins führt. In den meisten Fällen erfolgte diese Sequenzoptimierung jedoch unter gleichzeitiger Einbeziehung mehrerer Parameter und eine Abschätzung der Stärke des Einflusses der „codon usage“ ist schwierig. So wurde z.B. die Sequenz für Pfert aus *Plasmodium falciparum* in Richtung einer Senkung des AT-Gehaltes der kodierenden Sequenz (siehe auch 1.2), Vermeidung von Sekundärstrukturbildung der mRNA (siehe auch 1.3) und Codon-Anpassung an Hefe verändert. Dies führte zu einer Expression des Proteins sowohl in *S. cerevisiae* als auch in *P. pastoris* (Zhang et al. 2002). Für *P. pastoris* gibt es ebenfalls Arbeiten, in denen durch Änderung der „codon-usage“ eine Erhöhung der Expression erreicht werden konnte (Withers-Martinez et al. 1999; Outchkourov et al. 2002; Woo et al. 2002). In allen diesen Arbeiten wurde gleichzeitig der AT-Gehalt der Sequenzen verringert.

Die Gene für humanes Neurturin (Li et al. 2003) und für Lipase aus *Candida rugosa* (Brocca et al. 1998) wurden nur mit Blick auf eine „codon adaptation“ synthetisiert. Im Gegensatz zu den Wildtyp-Genen wird eine Expression in *P. pastoris* (humanes Neurturin) bzw. in *S. cerevisiae* und *P. pastoris* (Lipase) erzielt.

Die in *E. coli* vorkommenden Leserasterverschiebungen können ebenfalls durch Austausch der verursachenden Codone vermieden werden (Seetharam et al. 1988; Spanjaard und van Duin 1988; Kane et al. 1992; Calderone et al. 1996; Forman et al. 1998). Teilweise führten diese Mutationen zusätzlich zur korrekten Translation auch zu einer Erhöhung der Proteinausbeute (von 10 % des Gesamtproteins auf 30 %) und einem, im Vergleich zur Expression der Wildtyp-cDNA nicht mehr inhibierten Wachstum (Kane et al. 1992). Dies ist jedoch nicht immer der Fall. Ein Austausch von seltenen Codonen führt mitunter zur korrekten Translation, jedoch nicht zur Erhöhung der Proteinausbeute (Spanjaard und van Duin 1988).

Eine Möglichkeit, stille Mutationen in eine Sequenz einzuführen, ist die komplette Neusynthese von cDNAs, um sie somit an die „codon usage“ des Wirtes anzupassen: so wurde zur Expression von humanem TEL die kodierende Sequenz so verändert, dass eine Erhöhung des „codon adaptation index“ (CAI) (Sharp und Li 1987) resultierte. Der CAI ist ein Maß für die Anpassung einer gegebenen Sequenz an die optimale „codon usage“ eines Referenzorganismus, in diesem Fall wurde sie von 0,174 auf 0,754 erhöht, wobei der Index zwischen 0 und 1 variieren kann und 1 die ausschließliche Nutzung optimaler Codone bedeutet. Dies führte zu einer deutlichen Erhöhung der Expression auf ca. 20 % des Gesamtproteins (Martin et al. 1995).

Die komplette Neusynthese des Gens für humane Phosphatidylcholinesterase erhöhte den CAI von 0,25 auf 0,70. Eine Erhöhung der Expression von „gerade detektierbar“ auf 10 % des Gesamtzellproteins war die Folge (Feng et al. 2000). Analog wurde die Expression von Dihydrofolatreductase-Thymidylatsynthase aus *Plasmodium falciparum* um den Faktor zehn erhöht (Prapunwattana et al. 1996).

In zumindest einem Fall (Alexeyev und Winkler 1999) wurde durch ein synthetisches, codon-optimiertes Gen keine Erhöhung der Proteinexpression erreicht, in anderen Fällen wurde durch „codon adaptation“ sogar eine Verschlechterung der Expression beobachtet (Griswold et al. 2003). Als Ursache für die Verringerung der Expressionshöhe wird die Ausbildung von Sekundärstrukturen in der mRNA – siehe hierzu 1.3.1 – postuliert (Griswold et al. 2003).

Eine weitere Strategie, Limitationen durch seltene Codone auszugleichen, besteht in der Möglichkeit, die entsprechenden Anticodon-tRNAs überzuexprimieren.

So konnte der negative Einfluss der seltenen Arginin-Codone AGA und AGG auf die Expressionshöhe mittels Co-Überexpression des *E. coli* Gens *argU* – kodierend für die seltene tRNA_{AGA/AGG} – kompensiert werden (Acosta-Rivero et al. 2002). Hierbei wurde die Expression von drei heterologen Genen – jeweils reich an den genannten seltenen Codonen –

mit und ohne Überexpression von *argU* verglichen. Auf zwei der getesteten Sequenzen zeigte die Koexpression einen positiven – Faktor fünf bzw. drei –, auf die dritte Sequenz hingegen keinen Einfluss. Dies deutet auf inhibierende Einflüsse unabhängig von bzw. zusätzlich zu den ungünstigen Arginin-Codonen hin (Acosta-Rivero et al. 2002). Ein vergleichbarer Ansatz wurde mittels Überexpression von *ileX*, *argU* und *leuW* – kodierend für Anticodon tRNAs zu den seltenen Codonen ATA, AGA und CTA – verfolgt (Zdanovsky und Zdanovskaia 2000). Als Testproteine dienten sechs verschiedene Gene bzw. Genfragmente aus Clostridien. Der Effekt der drei tRNA-Gene auf die Expression reicht von „stark“ (*ileX*) über „moderat“ (*argU*) bis „nicht erkennbar“ (*leuW*). Es zeigte sich keine Korrelation zwischen der Wirkung der jeweiligen tRNA und der Häufigkeit der entsprechenden Codone. Interessanterweise war der Effekt der Koexpression von *argU* höher für ein Gen, in dem die entsprechenden zehn Codone gehäuft auftraten, als für eine Sequenz, bei der elf entsprechende Codone eher gleichmäßig verteilt auftraten.

Basierend auf diesen Ergebnissen wurden in jüngerer Zeit vermehrt Stämme mit zusätzlichen Kopien der tRNA-Gene für seltene Codone benutzt (Stratagene, BL21-CodonPlus® Serie) – (z.B. Senejani et al. 2001; Talarico et al. 2001) – hier sind aufgrund des parallelen Einsatzes mehrerer tRNA-Gene keine Einzeleinflüsse der jeweiligen Codone auseinander zuhalten. In den Untersuchungen war jedoch eine Expression des jeweiligen Zielproteins in Stämmen mit Wildtyp tRNA Ausstattung nicht detektierbar.

Leserasterverschiebungen durch seltene Codone können ebenfalls durch Überexpression der entsprechenden tRNA kompensiert werden (Calderone et al. 1996; Forman et al. 1998; McNulty et al. 2003).

Ko- bzw. Überexpression von tRNAs kann jedoch auch unerwünschte Effekte zeigen. So führte die Überexpression von tRNA^{Gly1} – spezifisch für GGG Codone – zu einer –1 Leserasterverschiebung an GGA Codonen (O'Connor 1998).

Eine andere Strategie, die auf das 5'-Ende der Sequenz abzielt, ist eine Veränderung des Proteins. So gelang die Expression des vesikulären Monoamin Transporters aus Ratte – eines Membranproteins – in *S. cerevisiae* nach der N-terminalen Fusion mit einer Abfolge von hefepräferierten Codonen. Die Höhe der Expression stieg mit der Länge dieses Bereiches an. Die Autoren postulieren als Ursache dieses Effektes eine verbesserte Translatierbarkeit bzw. eine Verbesserung der Translationsinitiation (Yelin und Schuldiner 2001).

Der Vollständigkeit halber soll erwähnt werden, dass die Kultivierungsbedingungen ebenfalls einen Einfluss auf die Fehleinbaurate bei Expression in *E. coli* haben. Bei der Expression der a1-Homeodomäne aus *S. cerevisiae* wurde ein Einfluss der Medienzusammensetzung auf die

Fehleinbaurate beobachtet (Forman et al. 1998). Diese Ergebnisse weisen auf einen verfahrenstechnischen Ansatz zur Lösung des Problems.

1.2 Der Einfluss der Nukleotidkomposition auf die Expressionshöhe

Für verschiedene Wirtsorganismen wurde die Nukleotidkomposition der kodierenden Sequenz als kritischer Parameter erkannt.

1.2.1 Der Einfluss der Nukleotidkomposition in *E. coli*

Während direkte Einflüsse der Nukleotidkomposition auf die Expression in *E. coli* als Expressionswirt nicht beschrieben wurden, gibt es hier jedoch Probleme bei sehr AT-reichen Sequenzen. MSP-1 aus *Plasmodium falciparum* ließ sich erst nach Senkung des hohen AT-Gehaltes (74 %) stabil in *E. coli* klonieren (Pan et al. 1999).

1.2.2 Einflüsse der Nukleotidkomposition auf die Expression in Hefe

In Hefe können Unterschiede in den Transkriptionsterminations- und Polyadenylierungssignalen von verschiedenen Organismen kritisch für eine heterologe Expression sein (Zhao et al. 1999). Zufällige Basenfolgen aus Genen höherer Eukaryonten können in Hefe als Transkriptionsterminations- und Polyadenylierungssignale erkannt und zu einer entsprechenden Prozessierung der mRNA führen. Allgemein sind die Terminations- bzw. Polyadenylierungssignale in Hefe weniger stark konserviert und variabler als in anderen Organismen, insbesondere Säugern. Sie bestehen aber – ähnlich wie die entsprechenden Säugersequenzen – aus AT-reichen Bereichen (Zhao et al. 1999) (siehe Abbildung 1).

Solche Effekte auf die heterologe Expression wurden sowohl für *S. cerevisiae* (Romanos et al. 1991; Milek et al. 2000) als auch für *P. pastoris* (Scorer et al. 1993; Gurkan und Ellar 2003) beschrieben. Wegen der schwach konservierten Sequenzen und der variablen Abstände bzw. Reihenfolge der Hefe-Polyadenylierungssignale konnten die kritischen Stellen in der Sequenz jedoch nicht eindeutig identifiziert werden (Romanos et al. 1991).

Ein grundsätzliches Problem bei der Betrachtung sequenzbasierter Merkmale wie z.B. „codon adaptation“ oder GC-Gehalt ist deren Abgrenzung gegeneinander. Verändert man eine der Größen, hat dies häufig Einfluss auf die jeweils andere. Den Versuch, den Einfluss der „codon-adaptation“ gegen den von AT-reichen Sequenzen in *P. pastoris* abzugrenzen machten (Sinclair und Choy 2002). Das Gen für humane Glucocerebrosidase wurde zunächst hinsichtlich der „codon usage“ optimiert, wodurch gleichzeitig der GC-Gehalt stieg. Die Expression wurde verglichen mit der eines Kontrollkonstruktes mit in gleichem Maße erhöhtem GC-Gehalt, aber unverändertem „Codon Adaption Index“ (CAI) (Sharp und Li

1987). Während das humane Gen zu einer sehr geringen Expression führte, wurde durch Codon- und GC-Gehalt Optimierung eine 10,6-fache und nur durch GC-Gehalt Erhöhung eine 7,5-fache Erhöhung der Proteinexpression beobachtet. Das identifiziert den GC-Gehalt als Haupteinflussgröße.

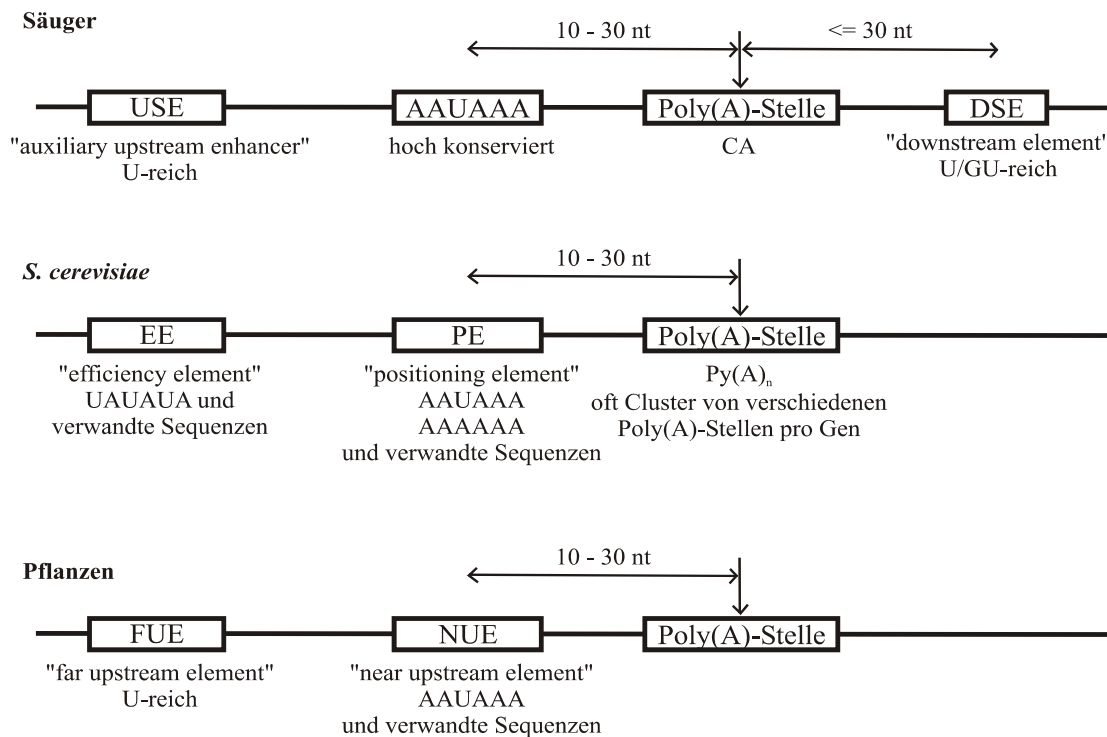


Abbildung 1: Vergleich von Transkriptionsterminationssignalen verschiedener Organismen (nach Zhao et al. 1990)

Die Abbildung zeigt die verschiedenen Konservierungsgrade bei unterschiedlichen Organismen. Man erkennt, dass die entsprechenden Sequenzen bei Säugern besser definiert sind als bei Hefe und Pflanzen. Zu erkennen ist ebenfalls der modulare Aufbau der Signale sowie die wenig konservierten Abstände der Module.

Interessanterweise wurden auch Unterschiede zwischen den beiden hauptsächlich als Expressionswirten genutzten Hefen *S. cerevisiae* und *P. pastoris* gefunden. Die Expression von HIV-1 ENV Protein in diesen beiden Wirten resultierte in einem kompletten Transkript in *S. cerevisiae*, in *P. pastoris* hingegen in einem vorzeitig terminierten 5'-Fragment (Scorer et al. 1993).

Die Hefe *Schwanniomyces occidentalis* hat relativ AT-reiche codierende Sequenzen (36 % GC-Gehalt, zum Vergleich *S. cerevisiae*: 40 % GC-Gehalt). Interessanterweise konnten in diesem System drei Gene (β -Lactamase, β -Glucoronidase und Chloramphenicoltransferase; alle aus *E. coli*), die sich in *S. cerevisiae* exprimieren lassen, nicht exprimiert werden (Janatova et al. 2003). Alle drei Gene zeichnen sich sowohl durch einen, im Vergleich zu *S.*

occidentalis Genen hohen, GC-Gehalt aus (47 – 62 % GC-Gehalt) als auch durch eine andere „codon usage“ als *S. occidentalis* (soweit für *S. occidentalis* bekannt). Hingegen lassen sich zwei Gene aus *Staphylococcus aureus* (ein Phleomycin-Resistenz Gen und eine Chloramphenicol Transferase), die sowohl einen höheren AT-Gehalt (35 % bzw. 27 % GC-Gehalt) als auch eine, an *S. occidentalis* besser angepasste „codon usage“ aufweisen, exprimieren.

1.2.3 Die Nukleotidkomposition in filamentösen Pilzen

Vorzeitige mRNA Prozessierung an Positionen mit AT-reichen Sequenzbereichen wurden ebenfalls bei der Expression von humaner α -Galaktosidase in *Aspergillus niger* und *A. nidulans* und beobachtet (Gouka et al. 1997). Interessanterweise wird dieses humane Gen in *P. pastoris* gut exprimiert (Chen et al. 2000).

1.2.4 Die Auswirkungen AT-reicher Bereiche in Säugerzellen

Eine andersgeartete Auswirkung von AT-reichen Sequenzbereichen bei heterologer Expression in Säugerzellen wurde bei der Expression von HIV Proteinen beobachtet. Hier wurden AT-reiche inhibitorische Sequenzen (INS) identifiziert (Schwartz et al. 1992; Schwartz et al. 1992; Reddy et al. 1995), die nicht zu einer vorzeitigen Transkriptionstermination, sondern zu einer verminderten Halbwertszeit des Transkriptes führen. Diese Sequenzen sind dabei nicht analog zu den in Eukaryonten bekannten AUUUA Elementen aus 3'-UTRs, die in diesem Bereich zu einer Destabilisierung des Transkriptes führen (Mikaelian et al. 1996). Falsche Prozessierung bei der Expression von Säugerproteinen – inklusive Proteine aus humanpathogenen Viren - in Säugerzellen sind aufgrund des verwandten Systems weder zu erwarten noch beschrieben.

1.2.5 Möglichkeiten zur Vermeidung negativer Einflüsse der Nukleotidkomposition

Eine häufig angewandte Strategie zur Reduzierung des AT-Gehaltes ist wiederum die Einführung stiller Mutationen. Sowohl für *S. cerevisiae* (Romanos et al. 1991; Milek et al. 2000) als auch für *P. pastoris* (Scorer et al. 1993; Gurkan und Ellar 2003) konnte so eine frühzeitige Transkriptionstermination vermieden werden. Hierfür wird der AT-Gehalt auf einen Anteil von ca. 50 % gesenkt (z.B. Tetanustoxin C in *S. cerevisiae*: hier wurde der AT-Gehalt von 71 % auf 53 % reduziert (Romanos et al. 1991)).

Häufig wurde bei Synthese bzw. Mutagenese von Genen für die Expression in *P. pastoris* neben der Senkung des AT-Gehaltes gleichzeitig die „codon usage“ der in Hefe angepasst sowie teilweise eine Bildung von theoretischen „stem-loop“ Strukturen innerhalb des

Transkriptes – siehe 1.3 – vermieden (z.B. Withers-Martinez et al. 1999; Outchkourov et al. 2002; Woo et al. 2002; Zhang et al. 2002). Durch diese kombinierten Ansätze ist eine Evaluierung des Einflusses der einzelnen Faktoren nicht möglich. Es ist allerdings prinzipiell schwierig, alle verschiedenen Parameter vollständig voneinander zu trennen. Veränderungen der Sequenz können z.B. immer auch zu Veränderungen in der Sekundärstruktur des Transkriptes – siehe 1.3 - führen.

Eine weitere Möglichkeit ist ein Wechsel des Wirtsorganismus: Tetanustoxin C, dessen Wildtypallel in *S. cerevisiae* zu verkürzten Transkripten führte (Romanos et al. 1991), ließ sich in *E. coli* sowohl als unmodifizierte Version (Halpern et al. 1990) als auch als Fusionsprotein mit Thioredoxin mit Erträgen von bis zu 40 mg / L Kultur (Ribas et al. 2000) exprimieren. Exprimiert werden konnte das Tetanustoxin-C-Wildtyp-Gen ebenfalls in Tabak-Chloroplasten (Tregoning et al. 2003). Hier wird sowohl das AT-reiche Wildtyp Allel als auch die AT-ärmere Version exprimiert.

Ein weiteres Beispiel ist die humane α -Galaktosidase, die sich nicht in *Aspergillus*, jedoch in *S. cerevisiae* und *P. pastoris* exprimieren lässt (Guisez et al. 1991; Chen und Shyu 1995; Gouka et al. 1997).

Organismen mit hohem AT-Gehalt der codierenden Sequenzen wie z.B. die Hefe *S. occidentalis* (36 % GC-Gehalt, Janatova et al. 2003) könnten sich als Wirte für die Expression AT-reicher Gene eignen. Ein Experiment mit AT-reichen – in anderen Hefen nicht exprimierbaren Sequenzen – könnte zeigen, ob *S. occidentalis* für die Expression AT-reicher Sequenzen besser geeignet ist als *S. cerevisiae* oder *P. pastoris*.

Die AT-reiche Sequenz von humaner α -Galaktosidase, die sich weder in *A. niger* noch in *A. nidulans* exprimieren ließ (Gouka et al. 1997), ließ sich z.B. in *P. pastoris* gut exprimieren (Chen et al. 2000). Möglicherweise sind filamentöse Pilze weniger zur Expression AT-reicher Sequenzen geeignet als Hefen.

Klonierungsprobleme aufgrund der Nukleotidkomposition konnten ebenfalls durch Senkung des AT-Gehaltes beseitigt werden. Die Reduzierung des AT-Gehaltes von MSP-1 aus *Plasmodium falciparum* von 74 % auf 55 % führte zur genetischen Stabilität des Konstruktes in *E. coli* und in der Folge auch zur Expression des Proteins (Pan et al. 1999).

Die mRNA-destabilisierende Wirkung von AT-reichen INS-Sequenzen bei der Expression in Säugerzellen konnte ebenfalls durch Einführung stiller Mutationen überwunden werden. Dies führte zu erhöhter Stabilität der mRNA und zu erhöhter Proteinausbeute (Mikaelian et al. 1996).

1.3 Hemmung der Translation: Einflüsse durch Sekundärstrukturen in der mRNA

Als ein weiterer wichtiger Parameter, der die Expression heterologer Proteine stark beeinflussen kann, ist die Sekundärstruktur der mRNA zu nennen. Studien hierzu konzentrierten sich auf relativ einfach einzugrenzende und theoretisch recht gut vorhersagbare sogenannte „stem-loop“ oder „hairpin“ Strukturen.

In verschiedenen Wirtsorganismen konnte für Einzelfälle ein Einfluss solcher Strukturen festgestellt werden. Die meisten Autoren konzentrierten sich hierbei auf Strukturen im Bereich der 5'-untranslatierten Region, des Start-Codons sowie bei *E. coli* im Bereich der Ribosomenbindungsstelle.

1.3.1 Die Folgen von Sekundärstrukturen der mRNA in *E. coli*

Im Bereich der 5'-UTR beeinflussen Sekundärstrukturen die Expression in *E. coli* negativ (Meetei und Rao 1998). Eine Sekundärstruktur, die innerhalb oder durch Teile der kodierenden Sequenz zustande kommt, hat ebenfalls einen negativen Einfluss auf die Expression (Bucheler et al. 1990; Bucheler et al. 1992; Humphreys et al. 2002; Griswold et al. 2003).

In zwei Arbeiten, in denen die Einflüsse von „codon adaptation“ und Sekundärstrukturen gegeneinander abgegrenzt wurden, dominierte der Einfluss der Sekundärstruktur über den seltener Codone in diesem Bereich (Wang et al. 1993; Griswold et al. 2003). Ein Codon-optimiertes Allel des Cutinase-Gens aus *Fusarium solani* wurde sogar geringer exprimiert als das Wildtyp-Allel, da es Sekundärstrukturen im 5'-Bereich aufwies (Griswold et al. 2003).

Verschiedene Autoren untersuchten den Zusammenhang zwischen thermodynamischer Stabilität der Sekundärstruktur und dem Grad der Expressionshemmung (de Smit und van Duin 1994; Griswold et al. 2003).

De Smit und van Duin (1994) fanden und quantifizierten eine solche Abhängigkeit, zum Teil unter Verwendung von Literaturdaten. Sie ermittelten bei Betrachtung von Experimenten zur Expression von vier verschiedenen Proteinen eine zehnfache Reduzierung der Expression bei einer Stabilisierung der Sekundärstruktur um $-1,4$ kcal/mol. Sie fanden einen unteren Grenzwert von einem ΔG_0 von -6 kcal/mol, eine instabilere Struktur zeigte keinen Effekt auf die Expressionshöhe. In einer anderen Arbeit wurde ein ΔG_0 von $-10,7$ kcal/mol als Grenzwert für eine heterologe Expression beschrieben (Bucheler et al. 1992). Eine mögliche Ursache der Abhängigkeit der Expressionshöhe von der thermodynamischen Stabilität liegt nicht notwendigerweise an der Fähigkeit des Ribosoms, eine nicht zugängliche Stelle zu

entwinden, sondern kann eine Folge des Lage des Gleichgewichtes, in der strukturierte und unstrukturierte Formen des Transkriptes vorliegen, sein (de Smit und van Duin 1994; de Smit und van Duin 2003). Dafür spricht auch die Beobachtung, dass durch eine Erhöhung der Komplementarität zwischen der jeweiligen Shine-Dalgarno Sequenz und des 3'-Endes der ribosomalen 16S RNA – also eine Verschiebung des Gleichgewichtes von der strukturierten mRNA zugunsten der rRNA Bindung - die Inhibierung der Expression durch mRNA Sekundärstrukturen in diesem Bereich teilweise aufgehoben werden kann (de Smit und van Duin 1994). Das postulierte Modell geht von einer unspezifisch an einen einzelsträngigen RNA-Bereich gebundenen 30S Untereinheit aus. Wenn die hemmende Sekundärstruktur in den einzelsträngigen Zustand wechselt, wandert diese zu der vorher blockierten Shine-Dalgarno Sequenz (de Smit und van Duin 2003).

Ein Teil der oben aufgeführten Faktoren wurde in ein mathematisches Model zur näherungsweisen Vorhersage der Expressionshöhe in *E. coli* umgesetzt (Ju et al. 1998). Vorhergesagt wird hier, ob das heterologe Protein mehr oder weniger als 20 % des Gesamtzellproteins ausmachen wird. Für die Quantifizierung der Parameter wurden die Expressionshöhen von 22 Genen – eingeteilt in zwei Klassen - herangezogen, von denen 13 über 20 % und die restlichen neun weniger als 20 % des Gesamtzellproteins ausmachten. In die relativ einfache Kalkulation gehen der „codon adaptation index“ (CAI) (Sharp und Li 1987) für die ersten 18 Basenpaare der kodierenden Sequenz sowie die freie Energie von mRNA Sekundärstrukturen im Bereich des Start- und des Stop-Codons ein. Als zusätzliche Rahmenbedingungen für eine hohe Expression müssen bestimmte Werte für den Abstand der Shine-Dalgarno Sequenz vom Start-Codon, ein unterer Grenzwert für die freie Energie einer eventuellen Sekundärstruktur im Start-Codon Bereich sowie ein oberer Grenzwert für eine eventuelle Sekundärstruktur im Bereich des Stop-Codons – d.h. eine instabile Struktur im 5'-Bereich, aber eine relativ starke im 3'-Bereich - erfüllt sein. Die Vorhersage wurde anhand eines Beispielproteins getestet: durch eine Anpassung der Sequenz für die humane Ricin-A Kette aus CD 28 im Sinne dieses Modells konnte die Expression von „nicht detektierbar“ auf 22 % des Gesamtzellproteins erhöht werden (Ju et al. 1998).

Der Versuch, die Wildtypversion des NS3 Protein des japanischen Encephalitis-Virus zu exprimieren, führte zu keiner detektierbaren Proteinmenge in *E. coli* (Satchidanandam und Shivashankar 1997). Die kodierende Sequenz bildet im Transkript eine „stem-loop“ Struktur, die mit dem Start-Codon beginnt. Die Einführung eines alternativen Start-Codons - 5' des eigentlichen Start-Codons liegend und resultierend in einem um vier Aminosäuren verlängerten Protein – führte zur Expression des selbigen. Dieses Resultat weist darauf hin,

dass eine Hemmung durch Sekundärstrukturmerkmale im Bereich der Translationsinitiation bzw. der Bindung des Ribosoms verursacht wird. Ein bereits translatierendes Ribosom scheint in der Lage zu sein, solche Strukturen aufzulösen.

Der Vollständigkeit halber soll an dieser Stelle erwähnt werden, dass in mehreren Arbeiten in Bezug auf die Translationsinitiation in *E. coli* ein starker Einfluss der Shine-Dalgarno Sequenz bzw. der Komplementarität selbiger zur rRNA festgestellt wurde (z.B. Wang et al. 1993; de Smit und van Duin 1994), dies ist jedoch kein Parameter, der direkt der kodierenden Sequenz zuzuordnen ist, sondern eher in die Konzeption des Expressionsvektors Eingang finden sollte.

Unter dem gleichen Gesichtspunkt ist ein weiterer Aspekt von Sekundärstrukturen in 5'-UTRs in *E. coli* zu nennen: diese können einen stabilisierenden Effekt auf die mRNA haben. Ist eine „stem-loop“ Struktur nahe genug am 5'-Ende des Transkriptes (nicht mehr als 2-4 Nukleotide entfernt), erhöht dies die Halbwertszeit des Transkriptes (Emory et al. 1992).

1.3.2 Der Einfluss von Sekundärstrukturen bei der Expression in Hefe

Für den Bereich der eukaryotischen Expressionssysteme ist der Einfluss von mRNA Sekundärstrukturen sowohl für die Expression von homo- als auch von heterologen Genen in der Hefe *S. cerevisiae* gut untersucht.

Ein translationsinhibierender Einfluss von Sekundärstrukturen in der 5'-UTR von *S. cerevisiae* Genen wurde in mehreren Fällen gezeigt (Baim und Sherman 1988; Cigan et al. 1988; Bettany et al. 1989; Oliveira et al. 1993; Vega Laso et al. 1993). Diese Inhibierung kann zu einer Reduktion der Expression auf „nicht mehr detektierbare Mengen“ führen. Ein schwacher Positionseffekt ist zu erkennen, da „stem-loops“ näher am Translationsstart geringfügig stärker hemmen als solche weiter in Richtung 5'-Ende des Transkriptes (Vega Laso et al. 1993). Allerdings reduzierte eine „stem-loop“ Struktur kurz vor dem Start-Codon die Expression eines Luziferase-Reporters immer noch auf bis zu 1,7 % der mit der Wildtyp 5'-UTR erzielten Aktivität (Niepel et al. 1999), die Expression von Chloramphenicolacetyltransferase (CAT) auf 2,03 % (Vega Laso et al. 1993).

Hierbei ist die Hemmung der Expression allerdings auch abhängig von der kodierenden Sequenz. Eine derzeit nicht berechenbare Wechselwirkung der Gensequenz mit der UTR war möglicherweise die Ursache für unterschiedliche Stabilitäten der Strukturen in der UTR (Oliveira et al. 1993).

„Stem-loop“ Strukturen, die das Start-Codon enthielten, führten ebenfalls zu einer starken Inhibierung der Expression (Oliveira et al. 1993).

Interessanterweise wird die Expression ebenfalls durch Strukturen unmittelbar nach dem Stop-Codon negativ beeinflusst. Dieser Effekt nimmt ab, wenn die Sekundärstruktur vom Stop-Codon in Richtung 3'-Ende des Transkriptes verlegt wird. Die Halbwertszeit der mRNA bleibt unbeeinflusst und ist nicht Ursache für den Effekt (Niepel et al. 1999).

Der Grad der Hemmung der Expression ist abhängig von der Stabilität der Sekundärstruktur (Vega Laso et al. 1993; Niepel et al. 1999). Am Beispiel bakterieller CAT wurde gezeigt, dass der Grad der Inhibierung mit der thermodynamischen Stabilität der Sekundärstruktur zunimmt. Stabile Sekundärstrukturen (-29,9 kcal/mol bzw. -49,4 kcal/mol), die das Start-Codon enthalten, resultieren in einer nicht mehr messbaren CAT-Aktivität (Vega Laso et al. 1993).

Das oben genannte Set von CAT-Reporter-Konstrukten mit Sekundärstrukturen unterschiedlicher Stabilität wurde hinsichtlich der Ribosomenbeladung der entsprechenden Transkripte näher untersucht (Sagliocco et al. 1993). Eine Auftrennung in einem Saccharose-Gradienten ergab für die unstrukturierten Transkripte einen einzelnen „peak“ mit einer gleichmäßig hohen Ribosomenbeladung der einzelnen mRNAs. Für die Transkripte mit stabiler Sekundärstruktur ergaben sich zwei „peaks“. Der Hauptanteil der Transkripte fand sich in der Region des Gradienten, die einer Beladung mit nur einem Ribosom entspricht, eine kleinere Fraktion enthielt die mRNAs, die mehrfach mit Ribosomen beladen sind. Die Autoren folgern, dass eine einmal entwundene Sekundärstruktur zur vollen Transkription und somit einer hohen Ribosomenbeladung des entsprechenden RNA-Moleküls führt – verursacht durch eine sterische Hinderung der Sekundärstrukturbildung durch „scannende“ ribosomale Untereinheiten (Sagliocco et al. 1993).

Der starke Einfluss von Strukturen nahe am oder um das Start-Codon bei Hefe (Oliveira et al. 1993; Vega Laso et al. 1993) ist nicht wie bei *E. coli* auf eine Hemmung der Rekrutierung des Ribosoms an die mRNA zurückzuführen, weil bei Eukaryonten die Rekrutierung des Ribosoms in der Regel über die „cap“-Struktur am 5'-Ende des Transkriptes erfolgt (Lewin 1990). Möglicherweise ist hier eine Inhibierung des „scanning“ oder der Translationsinitiation durch die Sekundärstruktur die Ursache für die verminderte Expression.

Die drastische Verringerung der Expression in Hefe durch eine stabile Sekundärstruktur nahe dem 5'-Ende der mRNA könnte durch eine Hemmung der Ribosomenrekrutierung verursacht sein (Vega Laso et al. 1993; Niepel et al. 1999).

Zusätzlich zu einer Inhibierung der Translation können Sekundärstrukturen auch zu einem Abbruch der Transkription im Bereich der „stem-loop“ Struktur führen. In den bis jetzt beobachteten Fällen wurde jedoch kein signifikanter Unterschied in der „steady-state“ Menge

des kompletten Transkriptes festgestellt und der Transkriptionsabbruch wurde nicht als Grund für die verminderte Expression angesehen (Vega Laso et al. 1993).

Komplementäre Sequenzen, die eine stabile Basenpaarung zwischen der 5'-UTR und der 3'-UTR ermöglichen, führten zu einer verringerten Halbwertszeit des Transkriptes von Chloramphenicolacetyltransferase und zu einer geringeren Expression (Vega Laso et al. 1993). Die Interaktion von 5'-UTR und 3'-UTR aufgrund von Basenpaarung wird ebenfalls als mögliche Ursache für reduzierte Expression von PGK postuliert (van den Heuvel et al. 1990).

1.3.3 Die Vermeidung von Expressionsproblemen aufgrund von Sekundärstrukturen in der mRNA

Durch ungerichtete stille Mutationen in dem Bereich, der für den N-Terminus der humanen Glutathionreduktase kodiert, konnte eine Erhöhung der Expression in *E. coli* um den Faktor 70 erreicht werden. Die höher exprimierenden Klone zeichneten sich durch einen reduzierten GC-Gehalt in diesem Bereich im Vergleich zum Wildtyp-Gen aus. Die Autoren vermuten als Ursache für die erhöhte Expression eine reduzierte Sekundärstrukturbildung (Bucheler et al. 1990).

Entsprechendes wurde, ebenfalls in *E. coli*, für die periplasmatische Expression von Fab-Fragmenten beobachtet (Humphreys et al. 2002). Hier führte eine ebenfalls ungerichtete Mutation der „wobble“-Positionen im Bereich des Signalpeptides zur verbesserten Expression. Selbiges konnte durch gezielte Einführung von „optimalen“ Codonen nicht erreicht werden. Postuliert wird eine Wechselwirkung der 5'-kodierenden Sequenz mit Promotorbereichen bzw. der 5'-UTR.

Eine Erhöhung der Expression von vier verschiedene pflanzlichen P450-Oxygenasen in *S. cerevisiae* wurde durch Reduzierung des GC-Gehaltes und somit Vermeidung stabiler Sekundärstrukturen sowie Adaption der „codon usage“ erreicht. Bei allen vier cDNAs konnte durch jeweils eine PCR-Reaktion mit einem 120 Nukleotide langen 5'-Primer der 5'-Bereich so verändert werden, dass eine Expression der Proteine resultierte (Batard et al. 2000). Der Einflusses der verschiedenen Faktoren wurde nicht getrennt erfasst, dieses Beispiel stellt allerdings eine relativ einfach durchzuführende Möglichkeit zur erfolgreichen Optimierung einer cDNA-Sequenz dar.

Ein Ansatz, der in einer Modifikation des Proteins resultiert, wurde für das NS3 Protein aus japanischem Enzephalitis-Virus durchgeführt. Hier wurde das 5'-Ende um vier Codone nach vorne verlagert, so dass die postulierte Sekundärstruktur hinter dem Start-Codon liegt. Diese

Veränderung führt zu einer Expression des Proteins (Satchidanandam und Shivashankar 1997).

Eine Maßnahme, die in den Bereich der Vektoroptimierung fällt, ist bei Expression in *E. coli* eine Erhöhung der Komplementarität zwischen der Shine-Dalgarno Sequenz und des 3'-Endes der ribosomalen 16S RNA (de Smit und van Duin 1994).

1.4 Sonstiges

Für Säugerzellen als Expressionswirt sind Studien zum Einfluss von heterologen Introns publiziert: durch die Insertion von zwei heterologen Introns in das intronfreie Gen für „green fluorescent protein“ konnte dessen Expression um den Faktor fünf erhöht werden (Lacy-Hulbert et al. 2001). Entsprechendes wurde für die Expression von Luziferase in verschiedenen Zellsystemen beobachtet (Xu et al. 2001).

1.5 Schlussfolgerungen

Zunächst ist zu sagen, dass Sequenzveränderungen meistens mehrere Parameter – z.B. „codon usage“ und GC-Gehalt – gleichzeitig beeinflussen. Aus diesem Grunde sind Schlussfolgerungen aus Experimenten, bei denen nur ein Parameter betrachtet wurde, schwer zu ziehen und es muss bedacht werden, dass möglicherweise Sekundäreffekte beobachtet worden sind.

Ein Einfluss der „codon usage“ auf die Expressionshöhe ist sowohl für *E. coli* als auch für eukaryotische Wirtssysteme beschrieben. Für *E. coli* ist, im Gegensatz zu anderen Organismen, ein Einfluss von seltenen Codonen im Gegensatz zu einer generellen „codon adaptation“ der gesamten Sequenz beschrieben. Für keinen der untersuchten Organismen gibt es Daten zu der Frage, ob es eine Proportionalität zwischen „codon adaptation“ und Expressionshöhe gibt, oder, falls es diese nicht gibt, welcher Grad von „codon adaptation“ einen Schwellenwert für eine Expression darstellt.

Interessant wäre das Experiment, ob Limitationen durch „codon usage“ in Eukaryonten ebenfalls durch Koexpression der limitierenden tRNAs wie in *E. coli* kompensiert werden können. Ein entsprechend modifizierter Wirtstamm wäre universell und würde Neusynthesen bzw. Mutationen der zu exprimierenden Sequenz überflüssig machen.

Speziell für 5'-Bereich der kodierenden Sequenz ist in Hefe ein Einfluss der „codon adaptation“ für den Fall einer Fusion eines langen Bereiches bevorzugter Codone an das 5'-Ende erkennbar (Yelin und Schuldiner 2001), wohingegen ein langer Bereich seltener Codone keinen Einfluss zeigt (Bettany et al. 1989). Für *E. coli* ist ein Fall bekannt, in dem die

Expression durch ein seltenes Codon im 5'-Bereich vermindert wird (Ma et al. 2003). All diesen Arbeiten ist gemeinsam, dass nicht festzustellen ist, ob die beobachteten Einflüsse spezifisch für das 5'-Ende der kodierenden Sequenz sind oder ob es sich um allgemeine Phänomene der „codon usage“ handelt.

Arbeiten, bei denen stille Mutationen ungerichtet, d.h. ohne die „codon adaptation“ zu verbessern, in das 5'-Ende der kodierenden Sequenz eingeführt wurden, führten sowohl für *E. coli* (Looman et al. 1987; Nilsson und Mannervik 2001; Stenstrom et al. 2001) als auch für *S. cerevisiae* (Looman et al. 1991) zu der Erkenntnis, dass die am besten exprimierten Allele keine optimale „codon usage“ in diesem Bereich aufweisen, also andere Ursachen die Expressionshöhe bestimmen. In diesem Zusammenhang sind die Effekte von mRNA-Sekundärstrukturen zu nennen. Für alle daraufhin untersuchten Wirtsorganismen konnte ein negativer Einfluss festgestellt werden. Bei einer Optimierung von Sequenzen sollte also zumindest im 5'-Bereich eher auf eine Vermeidung von Sekundärstrukturen als auf eine „codon adaptation“ optimiert werden.

In Pilzen ist eine frühzeitige Termination der Transkription aufgrund der Nukleotidkomposition ein wichtiger Faktor bei der heterologen Proteinexpression.

Eine Modifikation der Wirtszellen – analog zur Koexpression limitierender tRNAs – ist schwer vorstellbar. Eine gerichtete Mutation der kodierenden Sequenz ist hier ein Weg, zum Erfolg zu kommen.

Jedoch scheint dieser Einfluss in unterschiedlichen Wirten verschieden stark ausgeprägt zu sein. So gibt es ein Beispiel für eine Termination in *P. pastoris*, die nicht in *S. cerevisiae* auftritt (Scorer et al. 1993). Ebenfalls ein Beispiel gibt es für eine vorzeitige Terminierung in *Aspergillus* aber nicht in *P. pastoris* (Gouka et al. 1997; Chen et al. 2000). Mehr vergleichende Arbeiten würden zeigen, ob dies nur für diese Einzelfälle gilt oder allgemeine Phänomene sind. Ein Wechsel des Wirtsorganismus – falls unter Berücksichtigung von Faktoren wie z.B. Kultivierungsaufwand oder posttranslationalen Modifikationen sinnvoll – wäre eine Alternative zu einer Mutagenese der Sequenz.

In diesem Zusammenhang kann auch die gezielte Suche nach alternativen Wirtsorganismen, wie z.B. der Hefe *S. occidentalis* (Janatova et al. 2003) Lösungsansätze bieten.

Eine vollständige Transkription führt jedoch nicht immer zur Expression des Proteins. Ein Transkript der erwarteten Größe des AT-reichen Antigens Pfs48/45 aus *Plasmodium falciparum* in *S. cerevisiae* wird nach Senkung des AT-Gehaltes erreicht. Dies führt jedoch nicht zu einer Expression des Proteins (Milek et al. 2000). Dies zeigt, dass es verschiedene Einflüsse auf die heterologe Expression gibt, die gleichzeitig auftreten können. Das ist

angesichts des komplexen Ablaufes von Transkription, regulierter mRNA-Stabilität, Translation und Proteindegradation auch zu erwarten.

Die Tatsache, dass ein Einfluss des AT-Gehaltes auf die Transkription in *E. coli* nicht beschrieben ist, resultiert wahrscheinlich aus den unterschiedlichen Mechanismen der Transkriptionstermination in Pro- und Eukaryonten. In *E. coli* erfolgt diese auf zwei verschiedenen Wegen (zusammengefasst von Richardson 2002): die *Rho*-unabhängige Termination erfordert eine stabile – ca. 20 Basenpaare lange, GC-reiche - „stem-loop“-Struktur, gefolgt von sieben bis acht Adeninresten, *Rho*-abhängige Terminatoren sind weniger gut definiert, jedoch im allgemeinen sehr lang, bis zu 150 Basenpaare. Generell bindet *Rho* gut an poly(C) Bereiche. Diese definierten Anforderungen führen möglicherweise dazu, dass ein zufälliges Vorhandensein entsprechender Sequenzen in heterologen Genen selten ist.

Ein starker Einfluss von Sekundärstrukturen der mRNA ist bei allen daraufhin untersuchten Organismen zu erkennen (Bucheler et al. 1990; Bucheler et al. 1992; Vega Laso et al. 1993; Satchidanandam und Shivashankar 1997; Niepel et al. 1999; Humphreys et al. 2002). Für *E. coli* konnte sogar gezeigt werden, dass die Hemmung durch Sekundärstrukturen im Bereich des Start-Codons stärker ist als die seltener Codone (Wang et al. 1993; Griswold et al. 2003). In solchen Fällen muss die Gensequenz verändert werden, um zum Expressionserfolg zu kommen.

Alle diesbezüglichen Arbeiten betrachteten den Bereich um das Start-Codon. Analysen des restlichen Sequenzbereiches würden Aufschluss darüber geben, ob nur eine Hemmung von Ribosomenbindung bzw. Translationsinitiation entscheidend ist oder ob auch Sekundärstrukturen in anderen Bereichen der codierenden Sequenz ähnlich starke Folgen haben.

Eine weitere interessante Fragestellung ergibt sich aus der Beobachtung, dass sowohl in *E. coli* als auch in *S. cerevisiae* und anderen Eukaryonten endogene translatierte Sequenzen einen Trend zu mRNA Sekundärstrukturen zeigen (Seffens und Digby 1999; Katz und Burge 2003). Eine Erklärung für zugrundeliegenden Mechanismen gibt es nicht (Katz und Burge 2003). Diese Untersuchungen beziehen sich auf die theoretische freie Energie der Struktur des gesamten Transkriptes im Gegensatz zu den Untersuchungen zur heterologen Expression, die relativ stabile lokale Strukturen betrachten. Die Frage, ob eine stabile Faltung des gesamten Transkriptes eine heterologe Expression beeinflusst, ist noch für keinen Wirtsorganismus beantwortet.

Insgesamt ist zu sagen, dass es häufig kooperative Effekte verschiedener Parameter gibt (z.B. Milek et al. 2000; Sinclair und Choy 2002). Hier hilft die gleichzeitige Optimierung

verschiedener Einflussgrößen. Es gibt allerdings Parameter, deren Einfluss dominierend zu sein scheint. So kann in Hefe ein zu hoher AT-Gehalt bzw. ein zu hoher Gehalt an AT-reichen Bereichen zu gar keinem Protein führen (Romanos et al. 1991; Scorer et al. 1993). Der starke Einfluss von Sekundärstrukturen wurde bereits diskutiert. Eine Optimierung von Gensequenzen zur Expression sollte dies berücksichtigen. So wäre nach heutigem Kenntnisstand eine Vermeidung von stabilen Sekundärstrukturen am Beginn der kodierenden Sequenz wichtig, sowie eine Vermeidung AT-reicher Bereiche in eukaryotischen Expressionssystemen. Eine „codon adaptation“ sollte unter Berücksichtigung dieser beiden Parameter erfolgen. Letztere war nur dann erfolgreich, wenn massive Veränderungen der Sequenz erfolgen, z.B. die Vermeidung aller seltenen Codone oder eine Anhebung des CAI auf Werte von über 0,7.

Um mehr über die relativen Einflüsse der genannten Faktoren zu lernen, bzw. noch nicht berücksichtigte Parameter zu finden, sind weitere Untersuchungen nötig. Hier könnten z.B. verschiedene Kombinationen ungezielter stiller Mutationen über der gesamten Sequenz Aufschluss darüber geben, welche Sequenzmerkmale und in welchem Bereich der kodierenden Sequenz zur höchsten Expression führen.

Eine weitere Möglichkeit, die auch proteinbasierte Faktoren mit einschließt, wäre eine vergleichende Analyse verschiedener Sequenzen, die unterschiedliche Expressionshöhen zeigen. Dieses wurde in der nachfolgenden Arbeit versucht.

2 Problemstellung

Nach der Veröffentlichung der genomischen Sequenzen einer stetig wachsenden Zahl auch von eukaryotischen Organismen wie z.B. *S. cerevisiae* (Goffeau et al. 1996) oder *H. sapiens* (Lander et al. 2001) ist der nächste Schritt, die Funktionen der Gene und ihrer Produkte im genomischen Maßstab aufzuklären. Neben Untersuchungen des Transkriptoms (Okazaki et al. 2002), von Proteininteraktionen (Uetz et al. 2000; Gavin et al. 2002) oder Deletionsmutanten (Winzeler et al. 1999) ist die Aufklärung der dreidimensionalen Struktur der kodierten Proteine ein wesentlicher Beitrag dazu (Abbott 2000). Weltweit wurden Initiativen gegründet, um die Strukturen der Proteine für verschiedene Organismen in Hochdurchsatzansätzen zu lösen (z.B. Christendat et al. 2000; Goh et al. 2003; Terwilliger et al. 2003). Als eine von diesen hat die Proteinstrukturfabrik (PSF) in Berlin sich zum Ziel gesetzt, die Struktur humaner Proteine aufzuklären (Heinemann et al. 2000, www.proteinstrukturfabrik.de). Die vorliegende Arbeit ist innerhalb dieses Projektes angesiedelt.

Ein Flaschenhals bei der Strukturbestimmung durch Hochdurchsatzverfahren ist die heterologe Expression der Zielproteine (Vincentelli et al. 2003; Yokoyama 2003). Bestimmte Proteine lassen sich gar nicht oder nur mit geringen Ausbeuten heterolog exprimieren (z.B. Pikaart und Felsenfeld 1996; Milek et al. 2000; Boettner et al. 2002; Vincentelli et al. 2003).

Ein Ansatz, um eine möglichst große Zahl von Proteinen in ausreichender Menge zu produzieren, ist die Verwendung verschiedener Expressionssysteme. Aus diesem Grund sollte im Rahmen der PSF neben der Expression in *E. coli* und *S. cerevisiae*, *P. pastoris* als alternatives Expressionssystem etabliert werden. Die methylotrophe Hefe *P. pastoris* erlangte in den letzten Jahren wachsende Bedeutung als Expressionswirt (zusammengefasst von Faber et al. 1995; Cereghino und Cregg 2000), mit dem hohe Erträge an rekombinantem Protein erreicht werden können (Cereghino und Cregg 2000). Die Methoden für genetische Manipulation und Analyse sind ähnlich denen, die für *S. cerevisiae* gut etabliert sind. *P. pastoris* kann auf Minimalmedium zu sehr hohen Zelldichten wachsen und ist in der Lage, eukaryotische posttranslationale Modifikationen einzuführen. Die Verteilung und Länge von N-Glykosilierungen unterscheiden sich von denen in *S. cerevisiae*, insbesondere ist die Kettenlänge kürzer, was diese Hefe zu einer interessanten Alternative für die extrazelluläre Expression von Säugerproteinen macht (Grinna und Tschopp 1989). Mit dem Alkoholoxidase1 (AOX1) Promotor steht ein durch Methanol stark induzierbarer (Ellis et al. 1985) und durch Glukose und die meisten anderen C-Quellen strikt reprimierter (Tschopp et al. 1987) Promotor zur Verfügung, der das System zur Expression potentiell toxischer

Proteine sowie für Markierungsexperimente geeignet macht. Ein weiterer Vorteil von *P. pastoris* als Expressionssystem für strukturelle Untersuchungen ist die Möglichkeit der kostengünstigen C^{13} -Markierung der Proteine für kernresonanzspektroskopische Untersuchungen. Die Fütterung von C^{13} -markiertem Methanol ist eine preiswerte Alternative zu Glukose und führt zu Inkorporationsraten von 98 % (Laroche et al. 1994).

2.1 Entwicklung eines parallelisierten Systems zum „screening“ von Expressionsklonen

Im ersten Teil dieser Arbeit sollte ein parallelisiertes System zur Überprüfung von *P. pastoris* Klonen auf Proteinexpression entwickelt werden. Die cDNAs sollten in einem PSF-einheitlichen Standard kloniert und als Fusionsproteine mit N-terminalem His₆- und C-terminalem StrepII-tag exprimiert werden. Diese beiden Affinitätstags ermöglichen nicht nur eine immunologische Detektion, sondern auch eine chromatographische Reinigung der Proteine. Hierzu sollte zunächst ein PSF kompatibler Expressionsvektor konstruiert werden, die Klonierung der cDNAs sowie die Transformation von *P. pastoris* standardisiert und eine Überprüfung der Transformanten auf erfolgreiche Integration der Expressionskassette etabliert werden. Weiterhin sollte innerhalb der vorliegenden Arbeit ein System für das parallelisierte „screening“ der Hefeklone auf Proteinexpression im kleinen Maßstab entwickelt werden. Nach der Entwicklung sollte diese Methodik genutzt werden, eine Sammlung von charakterisierten Expressionsklonen für humane cDNAs zu erstellen.

2.2 Untersuchung der Sequenzen auf Parameter, die mit der beobachteten Expressionshöhe korrelieren

Im zweiten Teil der Arbeit sollte die zuvor erstellte Sammlung von charakterisierten Expressionsklonen genutzt werden, mittels bioinformatischer Anwendungen vergleichende Untersuchungen von cDNA-spezifischen Parametern, die die Expressionshöhe in *P. pastoris* beeinflussen könnten, durchzuführen. Die Sammlung von Klonen besteht aus Stämmen, die unter standardisierten Bedingungen auf Expression getestet wurden und sich lediglich in der zu exprimierenden cDNA unterscheiden. Die beobachteten Unterschiede in der Expressionshöhe müssen also durch diese bzw. durch das kodierte Protein bedingt sein.

Es sollte zunächst ohne eine *a priori* Annahme über mögliche Ursachen versucht werden, Sequenzparameter zu identifizieren, die mit der Expressionshöhe korrelieren. Ein solcher ungerichteter Ansatz erlaubt die Identifikation von unbekannten Einflussgrößen.

Weiterhin sollten auch bekannte Sequenzparameter wie z.B. „codon usage“ betrachtet werden, für die in Einzelfällen bereits ein Einfluss auf die Expressionshöhe beschrieben worden ist. Auf diese Weise sollte untersucht werden, ob diese Einflüsse generell auftreten. Darüber hinaus sollte die Verteilung von Parametern untersucht werden, für die bisher kein Einfluss auf die Expressionshöhe beschrieben ist, z.B. Hydrophobizität der Proteine oder isoelektrischer Punkt.

Experimentell sollten die Klone auf Unterschiede in mRNA-Menge und -Stabilität untersucht werden, um gegebenenfalls Zusammenhänge zu cDNA-abhängigen Parametern untersuchen zu können.

Die Analyse sollte auf intrazelluläre lösliche Proteine beschränkt bleiben, da die heterologe Expression von Membranproteinen zusätzlichen Limitationen unterliegt, die z.B. durch Transport bedingt sind (Butz et al. 2003). Diese Eingrenzung verhindert eine Zunahme der möglichen Faktoren auf einen bei gegebener Anzahl von Sequenzen nicht mehr sinnvollen Parameterraum.

Der Zusammenhang der Parameter mit der Expressionshöhe sollte statistisch beurteilt werden, um eine mögliche Signifikanz zu erkennen

3.1.4 Medien

LB: 1 % Trypton, 0,5 % Hefeextrakt, 0,5 % NaCl

für Platten: 1,5 % Agar

für Selektion auf Ampicillin-Resistenz wurde 100 µg / mL Ampicillin zugefügt

YNB: 0,67 % YNB w/o amino acids (Becton, Dickinson and Company, MD), 2 % Glukose

für Platten: 2 % Agar

YPD: 1 % Hefeextrakt, 2 % Pepton, 2 % Glukose

für Platten: 2 % Agar

Vitaminlösung für WM8/9 (pro 100 mL): Biotin 62,5 mg, Ca-Panthotenat 1,25 g, Nikotinsäure 250 mg, Pyridoxin 626 mg, Thiamin 250 mg; die Lösung wird sterilfiltriert und bei 4 °C gelagert.

Spurenelementlösung für WM8/9 (pro 100 mL): 0,4 mM CuSO₄, 1,8 mM FeSO₄, 0,5 mM MnCl₂, 0,4 mM Na₂MoO₄, 6mM ZnSO₄, 10 mM EDTA, sterilfiltriert

WM8: 10 g / L Na-Glutamat, 75 mg / L Inosit, 25 mg / L Mg₂Cl, 10 mg / L CaCl₂, 55 mg / L MgSO₄, 250 mg / L (NH₄)H₂PO₄, 2 g / L K₂HPO₄, 2,8 g / L NH₄Cl; nach dem Autoklavieren werden 40 mL / L 0,5 M Phosphatpuffer, pH 6,8, 1 mL / L Spurenelementlösung und 4 mL / L Vitaminlösung zugegeben; C-Quelle wird je nach Anwendung zugegeben.

WM9: 10 g / L Na-Glutamat, 75 mg / L Inosit, 25 mg / L Mg₂Cl, 10 mg / L CaCl₂, 55 mg / L MgSO₄; nach dem Autoklavieren werden 40 mL / L 0,5 M Phosphatpuffer, pH 6,8, 1 mL / L Spurenelementlösung und 4 mL / L Vitaminlösung zugegeben; C-Quelle wird je nach Anwendung zugegeben.

3.1.5 Puffer, Lösungen, Chemikalien

Agarose: SeaKem ME (Biozym, Hess. Oldendorf)

Blotpuffer: 25 mM Tris, 192 mM Glycin, 10 % Methanol

DNA-Größenmarker: „100-bp ladder“ (New England Biolabs, MA)

Ethidiumbromidlösung: Biorad, CA

„DIG DNA Labeling Mix“: (Roche Diagnostics, Mannheim)

Glassperlen: 0,25 – 0,5 µm (Roth, Karlsruhe)

HRP-Substrat: Western-Lightning™ Chemiluminescent Reagent Plus (Perkin Elmer, MA)

IMAC-Elutionspuffer: wie der Waschpuffer, aber mit 250 mM Imidazol

IMAC-Waschpuffer: 50 mM NaH₂PO₄, 300 mM NaCl, pH 8,0, 20 mM Imidazol

Maleinsäurepuffer: 0,1 M Maleinsäure, 0,15 M NaCl, pH 7,5 eingestellt mit NaOH

10fach MOPS-Puffer: 250 mM MOPS, 50 mM Na-Acetat, 20 mM EDTA;
pH 7,0 eingestellt mit NaOH

PBS-Puffer: 8,4 mM Na₂HPO₄, 1,6 mM KH₂PO₄, 150 mM NaCl

PMSF, Phenylmethylsulfonylfluorid: (Bethesda Research Laboratories, Bethesda, MA): 100 mM in Ethanol

RNA-Hybridisierungslösung: 5fach SSC-Puffer, 0,1 % (w/v) N-Lauroylsarcosin, 0,02 % (w/v) SDS, 1 % Boehringer Blockingreagenzlösung

RNA-Probenpuffer: 250 µl Formamid, 83 µl 35 %ige Formaldehydlösung, 50 µl 10*MOPS, 100 µl 50 % (v/v) Glycerin, 10 µl 2,5 % (w/v) Bromphenolblau

RNA-Waschpuffer: Maleinsäurepuffer mit 0,3 % (v/v) Tween 20

4fach SDS-PAGE Trenngelpuffer: 1,5 M Tris-HCl, 0,4 % (w/v) SDS, pH 8,8

4fach SDS-PAGE Sammelgelpuffer: 0,5 M Tris-HCl, 0,4 % (w/v) SDS, pH 6,8

SDS-PAGE Laufpuffer : 25 mM Tris, 192 mM Glycin, 0,1 % (w/v) SDS
4fach SDS-PAGE Probenpuffer nach Laemmli: 8 % (w/v) SDS, 20 % (v/v) Glycerin,
20 % (v/v) β -Mercaptoethanol, 0,25 M Tris-HCl pH 6,8, Bromphenolblau
20fach SSC-Puffer: 3 M NaCl, 0,3 M Na-Citrat, pH 7,0
StrepTactin-Elutionspuffer: wie der Waschpuffer, aber mit 5 mM Desthiobiotin
StrepTactin-HRP Konjugat: IBA, Göttingen
StrepTactin-Waschpuffer: 100 mM Tris-HCl, pH 8,0, 1 mM EDTA
„stripping“-Puffer: 625 mM Tris-HCl, pH 6,7, 2,0 % (w/v) SDS, es wurden vor
Gebrauch 14 μ l β -Mercaptoethanol pro 20 ml Puffer zugefügt.
TAE-Puffer: 20 mM Natriumacetat; 40 mM Tris; 2 mM EDTA; mit Eisessig pH 8,3
TEMED: Serva, Heidelberg
TFB I Lösung: 30 mM Kaliumacetat, 50 mM MnCl₂, 100 mM RbCl, 10 mM CaCl₂,
15 % (v/v) Glycerin, pH 5,8, sterilfiltriert
TFB II Lösung: 10 mM NaMOPS, 75 mM CaCl, 10 mM RbCl, 15 % (v/v) Glycerin,
pH 7,0, sterilfiltriert

3.1.6 Enzyme

Restriktionsenzyme wurden bezogen von New England Biolabs (MA) und entsprechend den
Herstellerangaben eingesetzt
Taq Polymerase (freundlicherweise zur Verfügung gestellt vom MPI für molekulare Genetik,
Berlin)
„proofreading“ Taq Polymerase (Qiagen, Hilden)
Alkalische Phosphatase (Roche, Mannheim) eingesetzt nach Herstellerangaben
T4 DNA-Ligase (New England Biolabs, MA)
S1-Nuklease (Bethesda Research Laboratories, MA) eingesetzt nach Herstellerangaben

3.1.7 Antikörper

anti-GFP-Antikörper (Santa Cruz Biotechnology, CA)
Penta-His-Antikörper (Qiagen, Hilden)
Sekundärer Antikörper; Schwein-anti-Hase; HRP-konjugiert (DAKO, Dänemark)
anti-DIG-Antikörper, HRP-konjugiert (Roche, Mannheim)

3.2 Programme, Server und Datenbanken

CodonW von John Peden; erhältlich im „worldwide web“ unter
www.molbiol.ox.ac.uk/cu/
DNASar DNASTAR Inc., WI Version 5.05
freak EMBOSS suite (Rice et al. 2000) Zugang über Institute Pasteur, Frankreich
(www.pasteur.fr)
GenBank (Benson et al. 2003) Zugang über www.ncbi.nlm.nih.gov/
MEME (Bailey 1994) Zugang über
<http://meme.sdsc.edu/meme/website/intro.html>
PESTfind (Rechsteiner und Rogers 1996); Zugang über
www.at.embnet.org/tools/bio/PESTfind/
PROSITE (Falquet et al. 2002) Zugang über <http://ca.expasy.org/prosite/>
PSIPRED (Jones 1999) Zugang über <http://bioinf.cs.ucl.ac.uk/psipred/>
(McGuffin et al. 2000).

SaccharomycesGenomeDatabase (SGD) (Dolinski et al.) (Stand: 02.06.2003)

SUPERFAMILY <http://supfam.org/SUPERFAMILY/hmm.html>
(Gough et al. 2001; Gough und Chothia 2002)

SwissProt (Boeckmann et al. 2003) <http://us.expasy.org/sprot/>

tblastn (Altschul et al. 1990) Nutzung über den Zentralrechner der PSF

3.3 DNA-Techniken

3.3.1 Plasmidisolation aus *E. coli*

Kleinere Mengen an Plasmid aus einzelnen Ansätzen (Mini-Präp, 1 – 5 µg Ausbeute) wurden mittels des GFX™Micro Plasmid Prep Kit (Amersham, UK) nach Herstellerangaben durchgeführt.

Mini-Präps im 96er Maßstab wurden mittels des Plasmid Miniprep₉₆ Kit (Millipore, MA) nach Herstellerangaben durchgeführt. Die notwendige Filtration erfolgte mittels Unterdruck. Größere Mengen an Plasmid (Midi-Präp) wurden mittels des Midi-Präp Kit (Qiagen, Hilden) ebenfalls nach Herstellerangaben gewonnen.

3.3.2 Restriktionsverdau

Alle Restriktionen wurden entsprechend den Empfehlungen der Enzymhersteller durchgeführt.

3.3.3 Reinigung von DNA-Fragmenten

Die Reinigung von DNA-Fragmenten (z.B. PCR-Produkte oder Restriktionsfragmente) erfolgte mittels des Plasmid Miniprep₉₆ Kit (Millipore, MA) nach Herstellerangaben. Die notwendige Filtration erfolgte mittels Unterdruck.

3.3.4 Ligation

Ungefähr 200 ng Plasmid und 150 ng PCR Produkt wurden zur Ligation zusammengegeben, Enzym und Puffer wurden entsprechend den Herstellerempfehlungen eingesetzt. Die Reaktion wurde für 1 – 2 Stunden bei Raumtemperatur durchgeführt.

3.3.5 Transformation von *E. coli*

Aus einer Gefrierkultur wurde eine 20 mL Vorkultur (LB-Medium) beimpft und über Nacht bei 37°C und 180 Upm inkubiert. 400 mL Hauptkultur wurden mit 4 mL der Vorkultur beimpft und über Nacht bis zu einer OD₆₀₀ von 0,4 – 0,5 inkubiert. Nach Abkühlen der Kultur auf Eis erfolgten alle weiteren Schritte bei 4°C. Die Zellen wurden geerntet und in 30 mL kalter TFB I Lösung resuspendiert. Nach Abzentrifugation der Zellen wurden diese in 4 mL kalter TFB II Lösung aufgenommen. Aliquots von 90 µl wurden bei –70°C gelagert.

Zur Transformation wurden 10 µl des Ligationsansatzes mit 80 µl kompetenten Zellen gemischt und 30 min auf Eis inkubiert, für 1 min bei 42 °C inkubiert und anschliessend mit 1 mL LB-Medium versetzt. Der Ansatz wurde 90 min bei 37 °C unter Schütteln inkubiert. Es wurden unterschiedliche Volumina auf LB-Platten mit Ampicillin ausplattiert und bei 37 °C über Nacht inkubiert.

3.3.6 Kolonie-PCR von *E. coli* Expressionsklonen zur Umklonierung der cDNAs

Die Amplifikation der cDNAs aus den *E. coli* Expressionsklonen des Teilprojektes 3 der PSF erfolgte mit den universellen Primern 5'-HisBamHI-f oder 5'-HisBglII-f (abhängig von der 5' Schnittstelle der klonierten cDNA, siehe 4.2.3.2) und 3'-NotIStrep-r (jeweils 0,2 µM). Es wurde eine Zahnstocherspitze Material in einen 20 µl PCR-Ansatz überführt. Die dNTPs wurden in einer Konzentration von 200 µM pro Nukleotid eingesetzt. Das Programm war 94 °C für 5 min, dann 30 Zyklen 94 °C 30 sek, 58 °C 30 sek, 72 °C 2 min, abschließend 72 °C für 10 min. Um Mutationen durch die PCR zu vermeiden, wurde „proofreading“ Polymerase benutzt.

3.3.7 Kolonie-PCR von *E. coli* zur Überprüfung der Klonierung

Von den selektierten Kolonien wurde eine Zahnstocherspitze Material in einen 20 µl PCR-Ansatz überführt. Die PCR erfolgte mit den universellen Primern 3'-AOX-r und 5'-AOX-f (jeweils 0,2 µM). Die dNTPs wurden in einer Konzentration von 200 µM pro Nukleotid eingesetzt. Das Programm war 94 °C für 5 min, dann 30 Zyklen 94 °C 15 sek, 55 °C 30 sek, 72 °C 30 sek, abschließend 72 °C für 10 min.

3.3.8 Transformation von *P. pastoris*

Die chemische Transformation von *P. pastoris* erfolgte nach einem Standardprotokoll (Invitrogen 1997).

Die Elektroporation von *P. pastoris* erfolgte mittels eines modifizierten Standard Protokolls (Invitrogen 1997). Die kompetenten Zellen wurden wie beschrieben hergestellt und in 80 µl Aliquots bei -70 °C gelagert. Die Plasmide zum Transformieren wurden, abhängig von der Sequenz des klonierten Inserts mit *SalI* oder *StuI* linearisiert und die Restriktionsansätze gegen H₂O für mind. 20 min dialysiert mittels Nitrocellulose Filter (0,025 µm; Millipore, MA). 3 – 5 µg linearisierter DNA wurden mit 40 µl kompetenter Zellen gemischt und für 5 min auf Eis inkubiert. Die Mischung wurde in eisgekühlte 0,2 cm Elektroporationsküvetten überführt. Die Parameter für die Elektroporation waren 1500 V, 50 µF und 200 Ω. Unmittelbar nach dem pulsen wurden die Zellen in 1 mL eiskalter 1 M Sorbitlösung suspendiert. Die Suspension wurde in zwei Aliquots auf YNB mit 2 % Glukose Platten ausplattiert und 3 – 4 Tage bei 28 °C inkubiert.

3.3.9 Kolonie-PCR von *P. pastoris*

Die stabile Integration der Expressionskassette wurde mittels Kolonie-PCR überprüft. Hierzu wurden selektierte His⁺-Kolonien zunächst in 100 µl YNB Medium mit 2 % Glukose überpickt und über Nacht bei 28 °C inkubiert. Dies war notwendig, um falsch positive Resultate zu vermeiden. 2 µl dieser Kultur wurden direkt in einen 20 µl PCR-Ansatz gegeben. Die PCR erfolgte mit den universellen Primern 3'-AOX-r und 5'-AOX-f (jeweils 0,2 µM). Die dNTPs wurden in einer Konzentration von 200 µM pro Nukleotid eingesetzt. Das Programm war 94 °C für 5 min, dann 30 Zyklen 94 °C 15 sek, 55 °C 30 sek, 72 °C 30 sek, abschließend 72 °C für 10 min.

3.4 Methoden der RNA-Analyse

3.4.1 Synthese und Markierung der Sonde

Die Sonde wurde per PCR von der 5'-untranslatierten Region des Vektors pPICHs synthetisiert (siehe auch Abb. 12). Es wurde 2,5 ng Plasmid eingesetzt. Die verwendeten

Primer waren AOX1UTR3'II und AOX1UTRkurz5' in 1 μ M Konzentration. Die dNTP's wurden in einer Konzentration von 100 μ M pro Nukleotid eingesetzt. Das markierte DIG-dUTP wurde mittels des „DIG DNA Labeling Mix“ eingeführt. Die Markierung der Sonde wurde per Agarosegel kontrolliert. Durch das eingebaute DIG läuft ein markiertes Fragment deutlich langsamer als ein unmarkiertes aus einer Kontrollreaktion mit unmarkiertem dNTP-Mix. Das PCR-Programm war: 94 °C 2 min; 30 Zyklen von 94 °C 15 sek, 40 °C 15 sek und 72 °C 30 sek, abschließend erfolgte ein Schritt von 72 °C für 10 min.

3.4.2 Isolation von Gesamt-RNA aus *P. pastoris*

Zur Isolation von Gesamt-RNA wurden je Probe 100 OD₆₀₀ geerntet. Die Pellets wurden einmal mit 1 mL kaltem 10 mM Tris/HCL pH 8,0 gewaschen. Anschließend können die Pellets bei – 70°C gelagert werden.

Zur Aufarbeitung wurden die Pellets in eiskalten 500 μ l 0,5 M NaCl/200 mM Tris/HCl pH 7,5/10 mM EDTA resuspendiert. Die Suspension wurde mit 500 μ l Glasperlen aufgefüllt. Es wurden 250 μ l auf 60 °C erwärmtes Phenol (nicht pH eingestellt!) und 25 μ l 10% SDS zugegeben. Nach drei min 100% vortexen wurden die Proben fünf min bei 60 °C inkubiert. Nach Abkühlung auf RT wurden 250 μ l Chloroform/Isoamylalkohol zugegeben. Nach kräftigem Schütteln wurde für drei min bei 13362 g zentrifugiert. Die obere Phase wurde mittels 500 μ l erwärmtem Phenol extrahiert. Die obere Phase wurde mittels Chloroform/Isoamylalkohol extrahiert und die RNA aus der oberen Phase mit 1/10 Volumen 3 M Na-Acetat, pH 4,8 und 2,5 Volumen Ethanol bei –70 °C für eine Stunde gefällt und dann 15 min bei 13362 g abzentrifugiert.

Die gefällte RNA wird in 30 μ l RNA-Probenpuffer gelöst.

3.4.3 RNA-Gel und Northern-blotting

Die in RNA-Probenpuffer gelösten Ribonukleinsäuren werden bei 65 °C für 10 min denaturiert und anschließend sofort auf Eis gestellt. Es wurden 20 μ l der Lösung auf ein Gel aufgetragen.

Die Gele waren 1,3 %ige Agarosegele in MOPS-Puffer versetzt mit 2 % (v/v) min. 35 %iger Formaldehydlösung und 1 μ l/mL Gelvolumen Ethidiumbromidlösung. Als Laufpuffer wurde 1*MOPS-Puffer benutzt, der Gellauf erfolgte bei 70 V bis der Bromphenolblaumarker ca. ¾ der Laufstrecke zurückgelegt hatte.

Das Gel wird zweimal für 15 min in 20fach SSC-Puffer gewaschen. Geblottet wurde mittels Kapillarblot über Nacht mit 20fach SSC-Puffer als Blotpuffer. Als Membran wurde Nylonmembran (Hybond, Amersham, UK) benutzt.

Nach dem Blotten wird die Membran zweimal kurz in 2fach SSC-Puffer gewaschen. Die RNA wurde auf der Membran durch Backen bei 80 °C für 2 h fixiert.

Die Vorhybridisierung erfolgte bei 38 °C für eine Stunde in Hybridisierungslösung unter leichtem Schwenken.

45 μ l der PCR-Reaktion zur Synthese der Sonde wurden bei 90 °C für 5 min denaturiert und anschließend sofort für 5 min auf Eis gestellt. Die denaturierte Sonde wurde in vorgewärmte (38 °C) Hybridisierungslösung gegeben. Die Hybridisierung erfolgte unter leichtem Schwenken bei 38 °C über Nacht. Ungebundene bzw. unspezifisch gebundene Sonde wurde durch zweimaliges Waschen bei RT für 10 min in 2fach SSC + 0,1 % (w/v) SDS gewegewaschen.

Die Nachweiß der Sonde erfolgte mittels HRP-konjugiertem anti-DIG-Antikörper (Roche, Mannheim). Alle hierzu nötigen Inkubationen erfolgten bei RT unter leichtem Schütteln. Zunächst wurde die Membran in RNA-Waschpuffer gewaschen. Danach wurde die Membran für 30 min in Maleinsäurepuffer + 1/10 Boehringer Blockinglösung blockiert. Zu dieser Lösung wurde der Antikörper zu einer Endverdünnung von 1:2000 zugegeben und 30 min

inkubiert. Danach wurde zweimal für 15 min mit RNA-Waschpuffer gewaschen. Die Detektion erfolgte mit HRP-Substrat nach Herstellerangaben. Visualisiert wurde mittels Videodokumentationssystem (Fujifilm LAS 100, Fuji Photo Film U.S.A., NY).

3.5 Methoden der Protein-Analyse

3.5.1 SDS-PAGE

Zur Auftrennung der Proteine wurde ein 12,5 %iges Trenngel und ein 4,5 %iges Sammelgel benutzt. Die Elektrophorese erfolgte mittels 0,75 mm dicker Gele in Mini-Protein[®] 3 Zellen (BioRad, CA). Die Größe der Trenngele betrug 8,5 * 5,5 cm. Der Lauf erfolgte bei 110 V bis das Bromphenolblau aus dem Gel austrat.

3.5.2 Western-Blotting

Der Transfer der Proteine erfolgte auf Immobilon PVDF-Membran (Millipore, MA). Der Aufbau des Blottes bestand aus je drei Lagen zugeschnittenem Whatman-Filter über und unter Gel und Membran. Die Filter wurden in Transferpuffer getrennt, die Membran mit Ethanol aktiviert und anschließend mit Transferpuffer benetzt. Es wurde ein semi-dry Blotter (Hölzel, Wörth) benutzt. Das Blotten erfolgte mit 500 mA pro Gel für 20 min.

3.5.3 Immunodetektion

Alle Inkubationen erfolgten unter leichtem Schütteln und, soweit nicht anders vermerkt, bei RT. Zunächst wurden die Membranen in PBS + 0.1 % (v/v) Tween 20 + 2 % BSA (w/v) entweder bei 4 °C über Nacht oder bei RT für eine Stunde blockiert. Zur Detektion des His₆-tags wurde dann 1 h mit Penta-His-Antikörper, verdünnt 1:2000 in PBS + 2 % (w/v) BSA inkubiert. Anschließend wurde dreimal mit PBS + 0.05 % (v/v) Tween 20 gewaschen. Die Inkubation mit dem sekundären Antikörper erfolgte ebenfalls für 1 h bei einer Verdünnung von 1:2000 in PBS + 0.05 % (v/v) Tween 20. Es folgten drei Waschschrte in PBS + 0.05 % (v/v) Tween 20. Die Detektion erfolgte mit HRP-Substrat nach Herstellerangaben. Die Visualisierung erfolgte entweder über Röntgenfilme oder mittels Videodokumentationssystem (Fujifilm LAS 100, Fuji Photo Film U.S.A., NY).

Routinemäßig wurde das Vorhandensein beider Tags überprüft. Vor der Detektion des StrepII-tags wurden die Membranen zunächst für 30 min bei 50 °C in „stripping“-Puffer inkubiert. Anschließend wurde zweimal in PBS + 0.1 % (v/v) Tween 20 gewaschen. Die Membranen wurden erneut wie oben blockiert. Um biotinylierte Proteine zu maskieren, wurde für 30 min in PBS + 0.05 % (v/v) Tween 20 + 2 µg / mL Avidin inkubiert. Zur dieser Lösung wurde StrepTactin-HRP Konjugat zu einer Endverdünnung von 1:4000 zugegeben und für 1 h inkubiert. Anschließend wurde dreimal mit PBS + 0.05 % (v/v) Tween 20 und dreimal mit PBS gewaschen. Die Visualisierung erfolgte wie oben.

3.5.4 Metall-Chelat Affinitätschromatographie

Die Kultivierungen erfolgten bei 28 °C und 180 Upm in 50 mL Maßstab in Schikanekolben. Die Vorkultur in WM9 mit 2,0 % (w/v) Glukose wurde drei Tage inkubiert, die Zellen anschließend mittels Zentrifugation bei 3345 g für 10 min geerntet und in frischem WM9 ohne C-Quelle resuspendiert. Die Induktion des Promotors erfolgte durch zweimal tägliche Zugabe von Methanol auf eine Endkonzentration von 1,0 % (v/v) und Glukose auf 0,1 % (w/v). Induziert wurde für drei bis vier Tage.

Die Ernte erfolgte in 10 mL Aliquots. Die Zellen wurden einmal mit 5 mL PBS gewaschen. Das gewaschene Pellet kann für einige Tage bei -70 °C gelagert werden.

Für den Aufschluss wurden die Zellen in 10 mL PBS + 0.05 % (v/v) Tween 20 + 1 mM PMSF resuspendiert. Nach Zugabe von 10 mL Glasperlen wurden die Zellen mittels 10 Zyklen von 1 min vortexen / 1 min auf Eis aufgeschlossen. Glasperlen und Zelltrümmer wurden mittels Zentrifugation bei 10000 g für 15 min abgetrennt.

Zum geklärten Zelllysat wurden 400 µl Ni-NTA Sepharose Matrix zugegeben und die Suspension bei 4 °C unter leichtem Schütteln inkubiert. Anschließend wurde die Suspension in eine leere Säule (Qiagen, Hilden) gegossen. Alle Chromatographieschritte erfolgten mittels Schwerkraftfluss. Die Matrix wurde dreimal mit je zwei Bettvolumen Waschpuffer gewaschen und anschließend das Protein mit dreimal je einem halben Bettvolumen Elutionspuffer eluiert.

3.5.5 StrepTactin Affinitätschromatographie

Vorkultur, Induktion der Expression und Lyse der Zellen erfolgte wie für die Metall-Chelat Affinitätschromatographie (siehe 3.5.4) beschrieben.

Eine leere Säule (Qiagen, Hilden) wurde mit 400 µl StrepTactin Sepharose (IBA, Göttingen) befüllt. Die Säule wurde mit drei Bettvolumen Waschpuffer gespült. Alle Chromatographieschritte erfolgten unter Schwerkraftfluss. Das geklärte Lysat wurde auf die Säule aufgegeben. Die Säule wurde fünfmal mit je einem halben Bettvolumen Waschpuffer gewaschen. Eluiert wurde mit fünfmal je 100 µl Elutionspuffer.

3.5.6 Kolonie-Blot von *P. pastoris*

Die Kolonien wurden auf PVDF Membranen angezogen, die auf YNB-Agarplatten mit 2 % Glukose als C-Quelle lagen. Die Induktion der Proteinexpression erfolgte durch Wechsel der Membranen auf YNB-Platten mit 0,1 % Glukose, auf denen vorher 400 µl Methanol ausgestrichen wurde. Nach vier Tagen Induktion erfolgte die Lyse der Zellen

3.6 Bioinformatische Methoden

3.6.1 Erstellung phylogenetischer Bäume

Phylogenetische Bäume sowie die zugrundeliegenden multiplen Alignments wurden mittels des Programms DNASTar erstellt. Die Alignments wurden mittels des Algorithmus ClustalW (Higgins und Sharp 1988) erzeugt. Es wurden sowohl DNA- als auch Proteinsequenzen verglichen. Hierbei wurden mehrere Parameter variiert. Die getesteten Kombinationen für Alignments von Proteinsequenzen waren wie folgt:

- 1.: Substitutionsmatrix: „structural“; „gap penalty“: 10, „gap length penalty“: 10
- 2.: Substitutionsmatrix: „structural“; „gap penalty“: 10, „gap length penalty“: 5
- 3.: Substitutionsmatrix: „PAM 250“; „gap penalty“: 10, „gap length penalty“: 10

Die Bedingungen 2 und 3 wurden sowohl mit den translatierten cDNA-Sequenzen als auch mit den Sequenzen inklusive der Tags durchgeführt.

Getestete Parameter für DNA-Alignments waren:

- 1.: Matrix: „identity“; „gap penalty“: 10, „gap length penalty“: 10

Die Untersuchung der DNA-Sequenzen wurde mit den ungetagten Sequenzen durchgeführt.

3.6.2 Ermittlung und Charakterisierung von Sequenzmotiven

Gemeinsame Motive in einem Sequenzdatensatz wurden mittels MEME ermittelt. Die Datensätze wurden im FASTA Format eingelesen. Die eingestellten Parameter waren wie folgt:

„any number of repetitions“; „maximum number of motifs“: 50; „maximum width“: 50; „minimum width“: 6

Der Grenzwert für den „E-value“ war 10000.

Um die Signifikanz der gefundenen Motive abzuschätzen, wurden die Analysen mit Sequenzen gleicher Zusammensetzung jedoch zufälliger Reihenfolge der Aminosäuren („shuffled“) wiederholt. Die „P-values“ der in diesem Set gefundenen Motive wurden als oberer Grenzwert für signifikante Motive betrachtet.

Die gefundenen Motive wurden mittels der PROSITE Datenbank auf bekannte biologische Bedeutung geprüft.

3.6.3 Quantifizierung AT-reicher Regionen

Es wurden Plots der Nukleotidkomposition entlang der Sequenz durch das Programm *freak* erstellt. Die Parameter wurden gewählt wie folgt: „residue letters“ wurden auf AT gestellt, „stepping value“ auf 1 und „window size“ auf 30. Um eine Quantifizierung AT-reicher Regionen im Gegensatz zu einem Durchschnittswert über die gesamte Sequenz zu erhalten, wurden die resultierenden Tabellen wie folgt prozessiert: Positionen, denen von *freak* ein Wert von 0,6 oder höher zugewiesen wurde, wurden als Teil einer AT-reichen Region betrachtet. Die Werte dieser Positionen wurden entlang der cDNA Sequenz aufaddiert.

3.6.4 Gesamt GC-Gehalt und GC-Gehalt an dritten synonymen Positionen (GC3s)

Der Gesamtgehalt an GC und der GC-Anteil an dritten synonymen Positionen wurden mittels des Programms *CodonW* ermittelt.

3.6.5 Messung der „codon usage“

Die „codon usage“ wurde gemessen mit dem Programm *CodonW*. Die kalkulierten Parameter waren die Anzahl effektiver Kodone (*N_c*) (Wright 1990) und die Anpassung an die „codon usage“ eines Referenzorganismus in Form des „codon adaptation index“ (CAI) (Sharp und Li 1987). Wegen eines mangels an genomischer Daten von *P. pastoris* wurde der CAI gegen einen Satz von hochexprimierten Genen aus *S. cerevisiae* (Sharp und Cowe 1991) gemessen.

3.6.6 Die Verteilung seltener Codone

Acht Codone werden in *S. cerevisiae* als selten angesehen (Zhang et al. 1991). Diese sind AGG (Arg); CGA (Arg); CGG (Arg); CGC (Arg); CCG (Pro); CUC (Leu); GCG (Ala); UCG (Ser). Es wurden sowohl die Frequenz dieser Codone pro cDNA als auch ihre absolute Anzahl bestimmt.

3.6.7 Generelle Proteinkennzahlen

Die berechneten Proteinkennzahlen sind der isoelektrische Punkt, die durchschnittliche Hydropathizität (GRAVY index), Aromatizität und Proteinslänge.

Proteinslänge und isoelektrischer Punkt wurden der PSF Datenbank entnommen. Letzterer wurde berechnet nach (Ribeiro und Sillero 1991). GRAVY index, der Mittelwert der Hydropathieindices der einzelnen Aminosäuren, und die Aromatizität, die Frequenz aromatischer Aminosäuren, wurden durch das Programm *CodonW* berechnet.

3.6.8 Proteindegradationssignale

3.6.8.1 PEST-Motive

PEST Motive (Rogers et al. 1986) wurden bestimmt mittels des Programms PESTfind bestimmt. Die „window size“ wurde auf 10 gesetzt. Die Ausgabewerte des Programms variieren zwischen -50 und 50. Für eine bessere Prozessierung dieser Daten wurden sie durch Addition von 50 in den positiven Bereich gebracht. Sequenzen, denen von PESTfind kein Wert zugewiesen wurde, weil kein PEST ähnliches Motiv gefunden wurde, wurde der Wert Null zugewiesen.

3.6.8.2 Lysinreste und die Abschätzung ihrer Oberflächenwahrscheinlichkeit

Lysinreste wurden wegen einer möglichen Ubiquitinierung und anschließenden Degradation der Proteine (Weissman 2001) als mögliche Determinanten für einen Abbau betrachtet. Die Oberflächenwahrscheinlichkeit für jeden Lysinrest wurde betrachtet, um die Zugänglichkeit des Restes für den Ubiquitinierungsapparat in Betracht zu ziehen. Die Oberflächenwahrscheinlichkeit wurde als Emini-Index (Emini et al. 1985) durch das Programm DNASTar berechnet. Eine hohe Oberflächenwahrscheinlichkeit wurde Lysinresten zugewiesen, die einen Emini-Index von über eins haben (Emini et al. 1985).

3.6.8.3 Hydrophobe Bereiche

Zur Ermittlung von hydrophoben Bereichen wurde mittels des Programms DNASTar ein Hydrophobizitätsplot nach Kyte-Doolittle (Kyte und Doolittle 1982) für jedes Protein erstellt. Aminosäurereste, denen eine Hydrophobizität von -1 oder weniger zugewiesen war, wurden als Teil eines hydrophoben Bereiches betrachtet. Die Werte dieser Reste wurden über die Sequenz aufaddiert.

3.6.9 Bestimmung der Ähnlichkeiten der untersuchten Sequenzen zu annotierten Hefeproteinen mittels BLAST

Die Quantifizierung der Ähnlichkeit der untersuchten Proteinsequenzen mit putativen *S. cerevisiae* Proteinen erfolgte über eine BLAST Analyse der Sequenzen gegen ein Set von putativen *S. cerevisiae* ORFs. Die Sequenzen von 6356 annotierten ORFs (Introns sind entfernt) wurden der *SaccharomycesGenomeDatabase* (SGD) entnommen. Da Proteinsequenzen mit Nukleotidsequenzen verglichen wurden, wurde das Programm tblastn benutzt. Als Schwellenwert wurde ein E-value von 10 gesetzt.

3.6.10 Zuordnung der Proteine zu strukturellen „protein superfamilies“, die auch in *S. cerevisiae* vertreten sind

Die Zuordnung zu „protein superfamilies“ erfolgte über die SUPERFAMILY Datenbank. Die Sequenzen wurden über die „worldwide web“ Eingabemaske im FASTA-Format eingegeben. Nach erfolgter Analyse wurde die Zuordnung der in den Sequenzen gefundenen Motive zum *S. cerevisiae* Genom („genome assignment“) betrachtet. In die Analyse einbezogen wurde, wie oft das entsprechende Motiv in *S. cerevisiae* vorkommt und in wie vielen verschiedenen Proteinen es vorhanden ist.

3.6.11 Vorhersage von Sekundärstrukturmerkmalen

Die Vorhersage von Sekundärstrukturmerkmalen (α -Helices, β -Faltblätter sowie Regionen, die nicht zu diesen Gruppen gehören („coils“)) wurde mittels PSIPRED durchgeführt.

3.7 Statistische Evaluierung

Aufgrund der geordneten Kategorisierung der Daten in vier Kategorien und der Tatsache, dass die Daten innerhalb der Kategorien nicht normalverteilt sind, wurde der verteilungsunabhängige Kruskal-Wallis Test angewandt (Kruskal und Wallis 1952). Hierdurch wird getestet, ob mindestens eine der Kategorien sich von den anderen signifikant unterscheidet. Es wurden alle Messdaten zunächst in einer Rangtabelle aufsteigend nach Messwert geordnet. Jedem Wert wurde eine Rangzahl entsprechend der Position in der Tabelle zugeordnet. Positionen in der Tabelle mit gleichem Wert bekamen die gleiche Rangzahl, die dem arithmetischen Mittel ihrer Position in der Tabelle entspricht. Nach der Ermittlung der Rangzahlen wurden die Werte in die Ausgangskategorien zurücksortiert und das Quadrat der Summe der Rangzahlen (R^2) gebildet. Die Prüfgröße **H** berechnet sich dann nach:

$$H = \left[\frac{12}{n(n+1)} \right] * \left[\sum_{i=1}^k \frac{R_i^2}{n_i} \right] - 3(n-1)$$

wobei **n** die Anzahl der Stichproben (hier: analysierte Sequenzen) ist und **k** die Anzahl der Kategorien.

Zur Durchführung des Testes wurde das Programm H-test in PERL erstellt (siehe Anhang). Das Programm liest eine aufsteigend geordnete Liste aller Werte, versehen mit einer laufenden Nummer und der Kategorie, und berechnet **H**.

Getestet wurde auf dem 5 % Niveau – d.h. eine Wahrscheinlichkeit von 95 %, dass mindestens eine der Kategorien einer anderen Grundgesamtheit als die anderen entstammt, wurde als signifikant angesehen. Hierfür muss **H** > 7,815 sein (Sachs 2002).

4 Ergebnisse

4.1 Optimierung der Expression im Schüttelkolben durch Induktions- und Medienvariation

Der angestrebte Kultivierungsmaßstab für die Charakterisierung der *P. pastoris* Expressionsklone war 2 mL in 24-„well“ Platten. In diesem Maßstab muss mit Limitationen der Zellen z.B. in Bezug auf die Sauerstoffversorgung bedingt durch eine schlechte Durchmischung gerechnet werden. Um ein möglichst niedriges Detektionslimit zu erzielen, wurde die Expression zunächst im 50 mL Maßstab optimiert. Außerdem stand somit für Experimente im Schüttelkolbenmaßstab (z.B. zur Isolierung kleinerer Mengen an rekombinantem Protein) ein optimales System zur Verfügung. Variiert wurden hierfür die verwendeten Medien und die Induktionsbedingungen des Promotors.

Die Optimierung erfolgte mittels des fluoreszenzphotometrisch zu quantifizierenden „green fluorescent protein“ (GFP) als Reporter.

4.1.1 Konstruktion und Evaluierung des Expressionsvektors für GFP

Zunächst wurde ein Expressionsvektor für GFP konstruiert. Aus dem Vektor pYEXTHSGFP (freundlicherweise zur Verfügung gestellt von O. Hesse) wurde mittels PCR die kodierende Sequenz für GFP inklusive eines N-terminalen His₆-tags und eines C-terminalen StrepII-tags amplifiziert. Das PCR-Produkt mit Überhängen wurde über *Bam*HI und *Eco*RI in den Vektor pPIC3.5 kloniert. Der resultierende Vektor pPICHGFPS wurde zur Integration in das *P. pastoris* Chromosom im *HIS4*-Gen linearisiert und alternativ mit *Bgl*II geschnitten. Die Linearisierung im *HIS4*-Gen führt bevorzugt zu einer Integration in den chromosomalen *His4*-Lokus über ein einzelnes Rekombinationsereignis, der Verdau mit *Bgl*II zu rekombinogenen Enden, die Promoter bzw. Terminator des *AOX1*-Lokus entsprechen. Letzteres setzt ein doppeltes Rekombinationsereignis für eine Integration voraus und führt zu einem Ersatz des chromosomalen *AOX1*-Lokus durch die Expressionskassette des Vektors.

Es wurde sowohl mittels Elektroporation als auch mit chemisch kompetenten Zellen transformiert. Bei Linearisierung des Vektors im *His4*-Gen wurden mittels Elektroporation typischerweise 5 – 10 Transformanten pro µg DNA, mit chemisch kompetenten Zellen 0 – 0,5 Transformanten pro µg Plasmid erzielt. Die Transformationsraten bei *Bgl*II-geschnittenem Vektor waren, bedingt durch das nötige doppelte Rekombinationsereignis, erheblich geringer. Aus diesen Gründen wurde im Hinblick auf den angestrebten Hochdurchsatz mit

einer Transformation mittels Elektroporation sowie im *His4*-Gen linearisierten Vektoren weitergearbeitet. Dies hat bezüglich der Methanol-Assimilation phänotypisch einen Wildtyp (Mut^+) zur Folge, da die endogene Kopie des *AOX1*-Gens erhalten bleibt.

Die Hefe-Transformanten wurden mittels Kolonie-PCR hinsichtlich der Integration der Expressionskassette getestet. Die Expression des Proteins wurde im 50 mL Maßstab mittels Western-Blot und Fluoreszenzmikroskopie überprüft (ohne Abbildung). Die Anzucht und Induktion erfolgte nach Invitrogen (1997).

4.1.2 Variation von Zufütterung und Medium

Bei der Benutzung des *AOX1*-Promotors für die Expression in *P. pastoris* ist Methanol der Induktor und die einzige C-Quelle. Um toxische Effekte durch zu hohe Methanolkonzentrationen zu vermeiden und gleichzeitig eine gute Induktion des Promotors zu erzielen, ist es wichtig, die Methanolkonzentration innerhalb eines engen Rahmens zu halten.

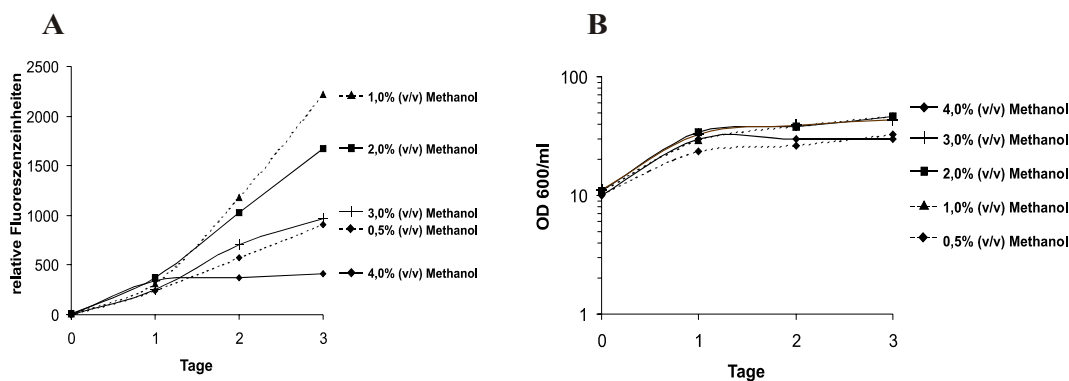


Abbildung 2: Vergleich verschiedener Methanolzugaben zur Induktion in WM9 Medium

Zur Optimierung der Induktion des *AOX1*-Promotors wurde die zugefütterte Methanolmenge variiert. Die Proteinexpression wurde über die Fluoreszenz des Reporterproteins GFP verfolgt. Wegen der Eigenfluoreszenz der Kulturen wurden jeweils die Differenzen zwischen einem GFP exprimierendem Stamm und einem Stamm, transformiert mit dem Kontrollvektor pPIC3.5, aufgetragen (relative Fluoreszenz). Die Wachstumsrate beider Stämme war praktisch gleich. Abb. 2A zeigt die GFP spezifische Fluoreszenz über der Zeit. Methanol wurde zweimal täglich zur angegebenen Endkonzentration zugegeben. Abb. 2B zeigt die Wachstumskurven des Expressionsstamms unter den verschiedenen Zufütterungsbedingungen. Alle Messungen wurden als Doppelbestimmungen durchgeführt.

Mittels zweier unabhängiger Mut^+ -Expressionsklone für „green fluorescent protein“ (GFP) wurde der Einfluss verschiedener Medien, die Zufütterung von Methanol sowie eine gemischte Zufütterung von Methanol und verschiedenen Konzentrationen von Glukose untersucht. Um die Expression zu quantifizieren, wurde die Fluoreszenz (Emission: 530 nm; Extinktion: 480 nm) eines 200 µl Aliquots der Kultur gemessen. Wegen einer Eigenfluoreszenz der Kultur wurde die gemessene Fluoreszenz um den Wert eines Kontrollstammes, transformiert mit dem Kontrollvektor pPIC3.5 ohne GFP-Gen korrigiert.

Der Kontrollstamm wurde in gleicher Weise kultiviert wie die Expressionsstämme und zeigte bei jedem Messpunkt eine vergleichbare Zelldichte. Die beiden Expressionsstämme zeigten bei allen Messpunkten vergleichbare Fluoreszenz. Im weiteren wird deshalb nur ein Stamm aufgeführt.

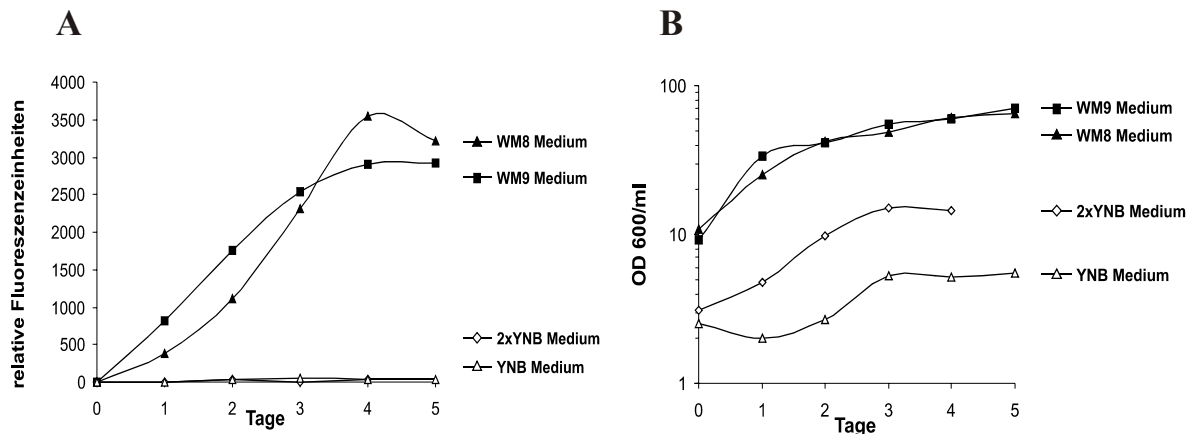


Abbildung 3: Vergleich der GFP-Expression in verschiedenen Medien

Die GFP-Expression und Wachstum in verschiedenen Medien während der Induktionsphase wurde verglichen. Aufgetragen wurden die Unterschiede zwischen einem GFP exprimierendem Stamm und einem Kontrollstamm, transformiert mit dem leeren Vektor. Abb. 3A zeigt die GFP spezifische Fluoreszenz über der Zeit in verschiedenen Medien. Die Induktion erfolgte durch Zugabe von Methanol zu einer Endkonzentration von 1,0 % (v/v) zweimal täglich. Abb. 3B zeigt die Wachstumskurven des Expressionsstammes während der Induktion. Die unterschiedlichen Start-OD's sind bedingt durch das unterschiedliche Wachstum der Vorkulturen, die in dem gleichen Medium wie die Hauptkultur geführt wurden. Die C-Quelle war in allen Vorkulturmedien 2,0 % (w/v) Glukose. Alle Messungen wurden als Doppelbestimmungen durchgeführt.

Variiert wurde zunächst die zugefütterte Methanolmenge. Die genannten Zugaben von Methanol beziehen sich auf die Endkonzentration im Medium und erfolgten zweimal anstatt einmal täglich wie im Standardprotokoll, um die Sprünge in der Methanolkonzentration geringer zu halten. Das benutzte Medium war WM9. Abbildung 2A zeigt, dass nach drei Tagen Induktionsphase eine Methanolzugabe von 1,0 % (v/v) zweimal täglich zu einer 2,5-fach höheren GFP spezifischen Fluoreszenz führt als eine Zugabe von 0,5 % (v/v). Eine höhere Zugabe von Methanol führt zu einer Abnahme der Fluoreszenz. Abbildung 2B zeigt die entsprechenden Wachstumskurven während der Induktionsphase. Das Zellwachstum variiert bei einer Methanolzugabe von 1,0 % (v/v) bis zu 3,0 % (v/v) nicht signifikant. Ein geringfügig reduziertes Wachstum erfolgt bei Zugabe von 0,5 % (v/v) und 4,0 % (v/v). Dies ist möglicherweise eine Folge von C-Quellen Limitierung (0,5 % (v/v) Methanol) bzw. eine Folge des toxischen Effektes von akkumulierendem Methanol in der Kultur (4,0 % (v/v) Methanol, Couderc und Baratti 1980). Die abnehmende Fluoreszenz bei Zufütterungsraten zwischen 2,0 % (v/v) und 3,0 % (v/v) Methanol ohne Einfluss auf das Zellwachstum deuten

auf eine optimale Konzentration zur Induktion des Promoters hin. Diese optimale Induktion durch Zugabe von 1,0 % (v/v) Methanol zweimal täglich wurde als Standard gesetzt.

Im folgenden wurden verschiedene Medien unter Verwendung dieser Zufütterungsstrategie geprüft. Abbildung 3A zeigt, dass die GFP-spezifische Fluoreszenz in WM8 und WM9 Medium um den Faktor 70 höher ist als sowohl in YNB als auch in zweifach konzentriertem YNB. Die Unterschiede in der Expressionsstärke spiegeln sich im Zellwachstum (Abbildung 3B) wieder.

Im Bioreaktor sind für *P. pastoris* gemischte Zufütterungsstrategien unter Verwendung von Glycerin und Methanol in der Expressionsphase entwickelt worden (Brierley et al. 1990; McGrew et al. 1997). Die Fragestellung war, ob vergleichbare Strategien auf den Schüttelkolbenmaßstab übertragbar sind.

In dieser Arbeit wurde Glukose anstelle des Glycerins verwandt. Verschiedene Konzentrationen von Glukose in Kombination mit 1,0 % (v/v) Methanol wurden getestet. Alle Kultivierungen erfolgten in WM9 Medium. Die genannten Konzentrationen sind Endkonzentrationen im Medium und die Zugabe erfolgte zweimal täglich.

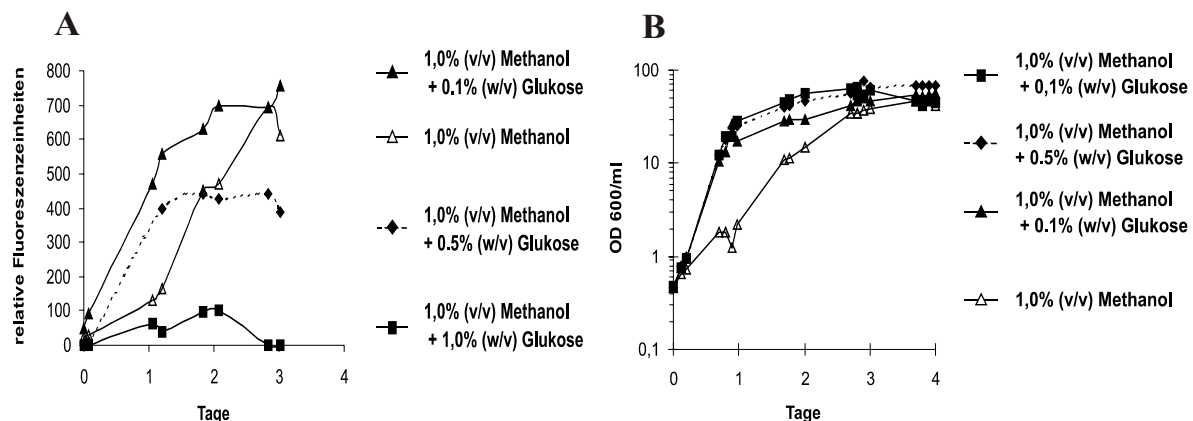


Abbildung 4: Gemischte Zufütterung von Methanol in Kombination mit verschiedenen Glukosekonzentrationen

Ermittelt wurden die Auswirkungen von verschiedenen Glukosemengen zusätzlich zu Methanol in der Zufütterung. Glukose und Methanol wurde zweimal täglich zu den angegebenen Endkonzentrationen zugefüttert. Abb. 4A zeigt die GFP spezifische Fluoreszenz eines Expressionsstammes über der Zeit. Abb. 4B zeigt die entsprechenden Wachstumskurven. Aufgetragen wurden die Werte für einen Expressionsstamm. Das Wachstum des Stammes transformiert mit dem Kontrollvektor war vergleichbar. Alle Messungen wurden als Doppelbestimmungen durchgeführt.

Eine Zugabe von 0,1 % (w/v) Glukose zusätzlich zu Methanol resultiert in einem schnelleren Anstieg der Fluoreszenz im Vergleich zu einer ausschließlichen Methanolzugabe (Abb. 4A). Die Höhe der Fluoreszenz nach einer Induktionszeit von drei Tagen ist für beide Zufütterungen gleich. Höhere Glukosekonzentrationen führen zu einer geringeren GFP-spezifischen Fluoreszenz. Dies ist bedingt durch die Glukoserepression des *AOX1*-Promotors

(Tschopp et al. 1987). Aus Abbildung 4B wird ersichtlich, dass die Zugabe von Glukose zu einer schnelleren Zunahme der Zelldichte führt. Das ist möglicherweise der Grund für die schnellere Zunahme der GFP-Menge.

4.2 Entwicklung eines parallelisierten Systems zum „screening“ von Expressionsklonen

Innerhalb der PSF wurde ein einheitliches Klonierungssystem etabliert. Die mit Schnittstellen versehenen Amplifikate der cDNAs können so innerhalb der PSF ausgetauscht werden und die cDNAs in die verschiedenen Expressionsvektoren kloniert werden. Über die Restriktionsschnittstellen *Bam*HI/*Not*I wird die cDNA gerichtet im Leserahmen kloniert. Im Rahmen der vorliegenden Arbeit wurde das *P. pastoris* Expressionssystem an den PSF-Standard angepasst und ein parallelisiertes System zum „screening“ auf Proteinexpression entwickelt.

4.2.1 Korrelation der Proteinexpression mit dem Nachweis der Expressionskassette mittels PCR

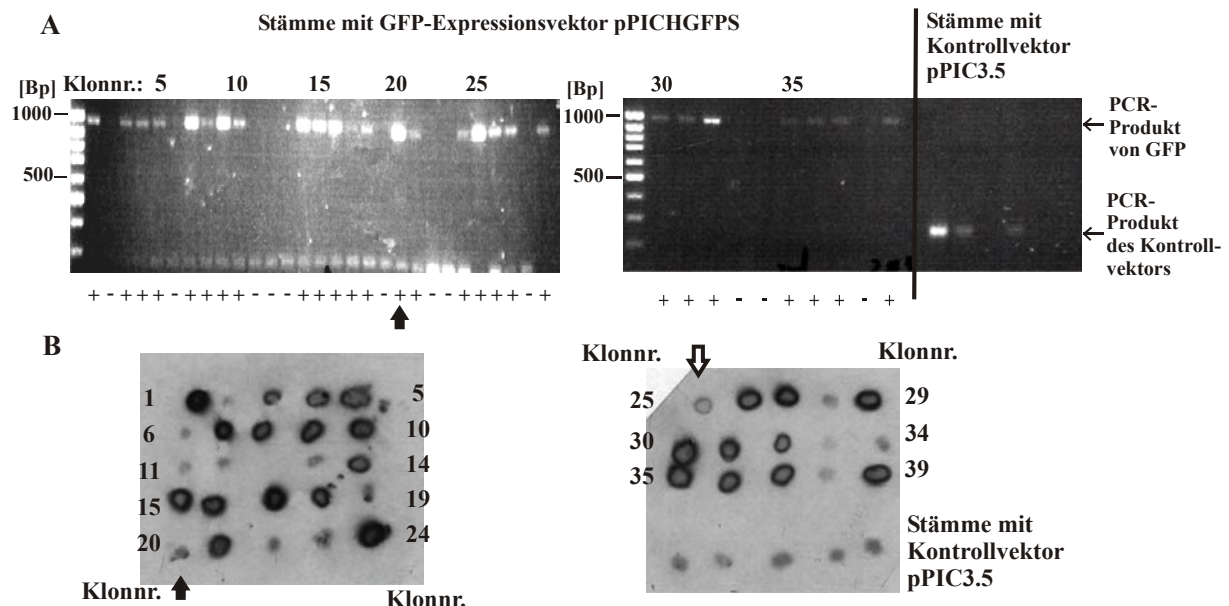


Abbildung 5: Korrelation des Nachweises der Expressionskassette mittels PCR (A) mit der Proteinexpression (B)

Teil A zeigt die Untersuchung der Transformanten auf das Vorhandensein der Expressionskassette mittels PCR. Die Analyse der Transformanten 1 – 39 ergab bei 27 Klonen ein PCR-Produkt der erwarteten Größe. Die Detektion des Proteins erfolgte mittels Kolonie-Blot und anti-GFP-Antikörper (Teil B). Klon Nr. 25 (⇔) zeigt im Kolonie-Blot ein relativ schwaches Signal, hier ergab ein Vergleich mit der Detektion unter schwachem UV, dass dieser Klon positiv für Proteinexpression war. Klon Nr. 20 (⇒) ist der einzige der auf DNA-Ebene positiven Klone, bei dem eine Expression von GFP nicht detektierbar war. Als Kontrolle wurden fünf Klone, die mit dem Kontrollvektor transformiert wurden, mitgeführt.

Zunächst musste die Frage beantwortet werden, wie viele Transformanten im Hinblick auf Proteinexpression getestet werden müssen, um keine cDNAs fälschlich aufgrund klonaler Variationen als „nicht exprimierbar“ einzustufen. In publizierten Protokollen wird auf die Notwendigkeit, Transformanten auf DNA-Ebene zu überprüfen, hingewiesen (Invitrogen 1997). Es waren allerdings keine Daten publiziert über die Anzahl der auf DNA-Ebene positiven Klone, die auf Proteinexpression hin überprüft werden müssen.

Zur Klärung dieser Frage wurden 39 individuelle His⁺ Kolonien getestet, die mit einem Expressionsvektor für GFP (pPICHGFPS) transformiert waren. GFP wird in *P. pastoris* gut exprimiert (siehe 4.1).

Die Kolonie-PCR der Klone ergab, dass 27 Kolonien (69 % der analysierten Klone) positiv in Bezug auf den Nachweis der Expressionskassette waren (Abb. 5A). Die Expression des Reporterproteins wurde mittels Kolonie-Blot unter Benutzung eines anti-GFP-Antikörpers (Abb. 5B) sowie über Detektion der Fluoreszenz der Kolonien unter schwachem UV-Licht untersucht. Von den 27 PCR-positiven Klonen zeigten 26 mit beiden Methoden eine Expression des Proteins. Das schwache Signal von Klon Nr. 25 im Kolonie-Blot konnte durch Vergleich mit der Fluoreszenz der Kolonien unter schwachem UV als positiv eingestuft werden. Klon Nr. 20 ist positiv in der PCR, zeigt aber keine detektierbare Proteinexpression. Die Ursachen hierfür sind nicht bekannt. Die 12 Klone, die in der PCR-Detektion negativ waren, waren auch im Kolonie-Blot negativ.

Zusammenfassend trugen von 39 His⁺-Klonen 27 die Expressionskassette, wovon 26 auch Proteinexpression zeigten. Hieraus wurde geschlossen, dass es ausreichend ist, zwei PCR-positive Klone auf ihre Proteinexpression zu testen. Eine Verifikation der Transformanten auf DNA-Ebene ist grundsätzlich notwendig, wie aus dem Anteil an Klonen, die sowohl in der PCR als auch im Kolonie-Blot negativ sind, hervorgeht.

Kontroll-Transformationen ohne Plasmid erzeugten unter Standard Bedingungen in keinem Fall His⁺-Kolonien. Dies zeigt, dass der Anteil an His⁺-Kolonien ohne Expressionskassette nicht durch Rückmutationen oder ungünstige Selektionsbedingungen nach der Transformation bedingt ist.

4.2.2 Konstruktion des PSF-Expressionsvektors

Der Vektor pPIC3.5 wurde modifiziert und dem Standard der PSF angepasst. Die Modifizierungen umfassen die Einführung einer Translationsinitiationssequenz (AAAATGTCT) gefolgt von einer Sequenz kodierend für sechs Histidinreste (His₆-tag). Weiterhin kodiert der Vektor für einen C-terminalen StrepII-tag gefolgt von einem Stop-

Codon (Abb. 6). Die PCR-amplifizierten cDNAs werden ohne eigenes Start- und Stop-Codon in die *Bam*HI und *Not*I Schnittstellen kloniert.

Für die Konstruktion des Vektors musste zunächst eine vorhandene *Not*I-Schnittstelle aus pPIC3.5 entfernt werden. Dies erfolgte über eine Linearisierung des Vektors durch *Not*I-Verdau und anschließender Entfernung der überhängenden Einzelstränge durch S1-Nuklease. Eine Selbstligation führte zum Vektor pPIC3Δ*Not*I. Die PSF-MCS inklusive der Tags wurde durch Ligation eines doppelsträngigen DNA-Fragments eingefügt.

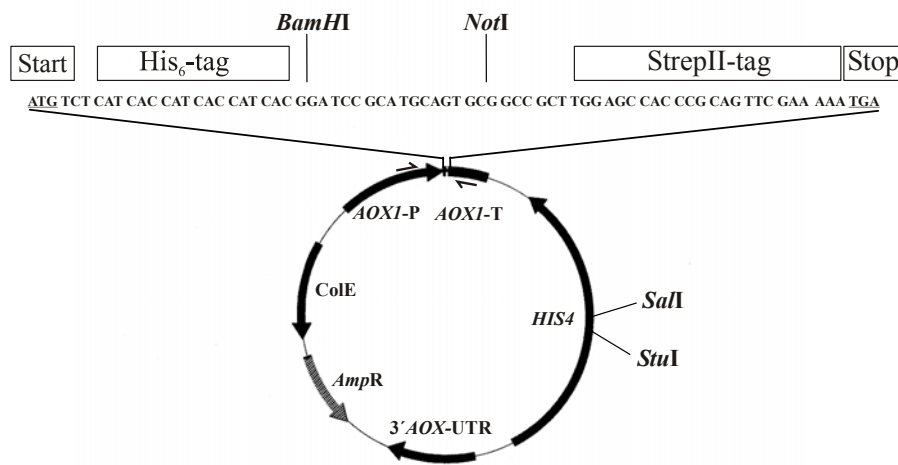


Abbildung 6: Das Expressionsplasmid pPICHs

Die Abbildung zeigt den Expressionsvektor pPICHs. Herausgehoben ist die DNA-Sequenz der Expressionskassette. Die Hybridisierungsstellen der Primer 5'-AOX1-f im *AOX1*-Promotor bzw. 3'-AOX1-r im *AOX1*-Terminator sind durch schwarze Pfeile markiert.

Dieses wurde durch Hybridisierung zweier Oligonukleotide (freundlicherweise zur Verfügung gestellt von Dr. C. Holz) erzeugt. Das Fragment enthält die Sequenzen für die beiden Affinitätstags sowie die Schnittstellen für eine *Bam*HI/*Not*I Klonierung. Am 5'-Ende befindet sich ein *Bam*HI kompatibler Überhang, am 3'-Ende ein *Eco*RI Überhang. Erstgenannter enthält nicht die 5'-Sequenz für eine Erkennung durch *Bam*HI und führt somit nach Ligation zu einer Eliminierung der *Bam*HI-Schnittstelle. Der Vektor pPIC3Δ*Not*I wurde mit *Eco*RI und *Bam*HI geschnitten und das Fragment einligiert. Die resultierende Expressionskassette des neuen Vektors pPICHs wurde durch DNA-Sequenzierung überprüft.

4.2.3 Klonierung und Expression verschiedener humaner cDNAs

Im folgenden wurden verschiedene humane cDNAs in den Expressionsvektor pPICHs kloniert, die resultierenden Konstrukte in *P. pastoris* transformiert und die entstandenen Hefeklone auf Proteinexpression getestet.

4.2.3.1 Auswahl der bearbeiteten cDNAs

Die Auswahl der bearbeiteten cDNAs erfolgte durch das Teilprojekt Bioinformatik der PSF (Gruppe von Dr. P. Bork am EMBL, Heidelberg, beschrieben in Holz et al. 2003). An alle humanen Proteine in GenBank (Benson et al. 2003) wurden folgende Auswahlkriterien angelegt: Proteine kleiner als 500 Aminosäuren, keine bekannte Struktur, keine Transmembrandomänen, keine „coiled-coil“ Regionen und keine Sequenzen geringer Komplexität. Nachdem redundante Sequenzen aus dem so erhaltenen Satz an humanen Sequenzen entfernt worden waren, wurde eine „expressed sequence tag“-Datenbank bestehend aus allen Klonen des I.M.A.G.E.-Konsortiums (Lennon et al. 1996) mittels BLAST nach cDNA-Klonen mit vollständiger Sequenz durchsucht. Von der Bioinformatik-Gruppe der PSF wurden 1212 cDNAs zur Bearbeitung vorgeschlagen.

4.2.3.2 Klonierung der cDNAs

Das Ablaufschema von der Klonierung der cDNAs bis zur Expressionskontrolle der resultierenden Hefeklone ist dargestellt in Abbildung 7. Die cDNAs, die für *P. pastoris* bearbeitet wurden, lagen zum Teil in Expressionsvektoren für *E. coli* (freundlicherweise zur Verfügung gestellt von Dr. V. Sievert und Dr. K. Büsow) oder *S. cerevisiae* (freundlicherweise zur Verfügung gestellt von Dr. C. Holz und N. Bolotina), zum Teil als PCR-Produkt (freundlicherweise zur Verfügung gestellt von Dr. C. Holz und N. Bolotina) vor.

Die erste Amplifikation aus den I.M.A.G.E.-Klonen zur Klonierung in die *E. coli* oder *S. cerevisiae* Expressionsvektoren erfolgte mit genspezifischen Primern, die zusätzlich mit den Sequenzen für die erforderlichen Schnittstellen (*Bam*HI/*Bg*II bzw. *Not*I) versehen waren. Eine *Bg*II-Schnittstelle wurde in den Fällen eingeführt, in denen die cDNA eine *Bam*HI-Erkennungssequenz enthielt. Der Entwurf der Primer resultierte in einer Klonierung der cDNAs im korrekten Leserahmen in die PSF-einheitliche Klonierungsstelle (Holz et al. 2003). Soweit keine PCR-Produkte vorhanden waren, wurden die cDNAs für die Umklonierung in den *P. pastoris* Expressionsvektor aus den *E. coli* Expressionsklonen mit universellen Primern amplifiziert (siehe 3.3.6). Alle PCR-Reaktionen erfolgten mittels „proofreading“-Polymerase, um Mutationen durch die PCR zu vermeiden.

Der Erfolg der PCR und die Größe der Amplifikate wurden auf einem 1,5 %igen Agarosegel überprüft.

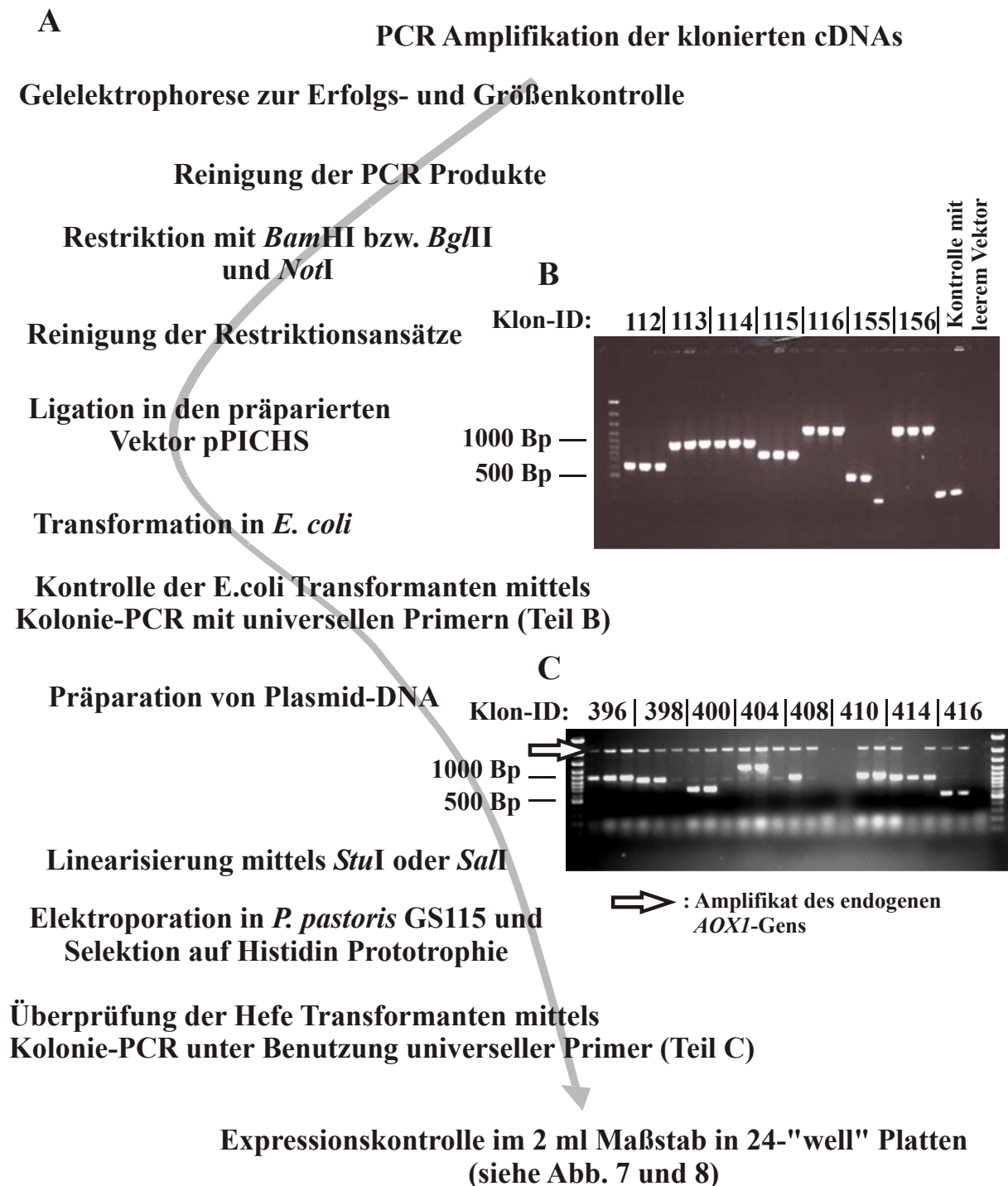


Abbildung 7: Arbeitsablauf der Klonierung, Transformation und Expressionskontrolle

Aufgezählt werden die einzelnen Arbeitsschritte von der Amplifikation der cDNAs bis zur Expressionskontrolle der resultierenden *P. pastoris* Klone (Teil A). Teil B zeigt ein Agarosegel einer *E. coli* Kolonie-PCR zur Überprüfung der Klonierung der cDNAs im *P. pastoris* Expressionsvektor. Überprüft wurden jeweils drei *E. coli* Klone pro cDNA. Plasmid-DNA zur Hefetransformation wurde von je einem Klon mit Insert in der erwarteten Größe isoliert. Teil C zeigt ein Gel einer Kolonie-PCR von *P. pastoris*. Es wurden jeweils drei Transformanten einer cDNA auf Integration der Expressionskassette überprüft. Da die verwendeten Primer (5'-AOX1-f und 3'-AOX1-r) im *AOX1*-Promotor bzw. Terminator hybridisieren, wird das endogene *AOX1*-Gen gleichzeitig amplifiziert (⇒). Für jede cDNA wurden zwei Transformanten mit nachgewiesener Expressionskassette auf Proteinexpression getestet. Teilweise wurde diese Zahl bei der Überprüfung von zwei Kolonien nicht erreicht (siehe Klon 408 in Teil C), in solchen Fällen wurden weitere Transformanten überprüft.

Von 114 ausgewählten cDNAs konnten alle amplifiziert werden. Von den PCR-Produkten hatten 15 eine Größe, die nicht mit den Angaben aus GenBank übereinstimmte. Dies kann zum einen durch unspezifische Amplifikate bedingt sein, zum anderen durch alternative Spleißprodukte. Da in jedem Fall von einer unbekannten Sequenz ausgegangen werden muss, wurde hier von einer weiteren Bearbeitung abgesehen. Der Vektor wurde durch sequentielles Schneiden mit *Bam*HI und *Not*I präpariert. Nach einer Reinigung der DNA erfolgte eine Dephosphorylierung mittels alkalischer Phosphatase. Der behandelte Vektor wurde erneut aufgereinigt.

Die PCR-Produkte wurden ebenfalls aufgereinigt und mit *Bam*HI bzw. *Bgl*II und *Not*I geschnitten und mit dem vorbehandelten Vektor ligiert.

Die Ligationen wurden in *E. coli* JM109 oder XL1-blue Zellen transformiert. Drei bis sechs Transformanten wurden auf das Vorhandensein eines Inserts mittels Kolonie-PCR überprüft (siehe Abb. 7B). Die benutzten Primer (5'-AOX1-f und 3'-AOX1-r) sind universell und hybridisieren am Promotor bzw. Terminator. Es wurden alle 99 cDNAs kloniert und pro cDNA wurde von einem Klon mit Insert in der erwarteten Größe Plasmid-DNA isoliert.

4.2.3.3 Transformation der Plasmide in *P. pastoris* und Überprüfung der Transformanten

Die präparierten Plasmide wurden mittels *Sal*I oder *Stu*I (abhängig von der Sequenz der klonierten cDNA) im *HIS4*-Gen linearisiert. Die Transformation in *P. pastoris* erfolgte mittels Elektroporation. Von den auf Histidin-Prototrophie selektierten Transformanten wurden mindestens drei mittels Kolonie-PCR auf die Integration der Expressionskassette getestet (siehe Abb. 7C). Es wurden immer zwei unabhängige, in der PCR-positive Transformanten auf Proteinexpression untersucht.

Für alle 99 Inserts konnten je zwei *P. pastoris* Klone mit integrierter Expressionskassette isoliert werden.

4.2.3.4 Entwicklung der Expressionskontrolle im 2 mL Maßstab

Um die Expressionskontrolle im kleinen Maßstab zu entwickeln und zu optimieren, wurde ein Klon gewählt, der das humane Homolog des Hefe Med7-Proteins (hMed7) exprimiert (interne ID 283, siehe Tabelle 1 im Anhang). Von acht verschiedenen Proteinen, deren Expression bis zu diesem Zeitpunkt im 50 mL Maßstab getestet wurde, wies dieses die geringste Expressionshöhe auf. Die Abschätzung der relativen Expressionshöhe erfolgte über einen Western-Blot. Die absolute Expressionshöhe wurde nach affinitätschromatographischer Reinigung und Coomassie-gefärbtem Gel (siehe Abb. 9) auf 2 – 5 mg heterologes Protein / L

Kulturvolumen quantifiziert. Da die erforderliche Menge an reinem Protein für Kristallisationsexperimente in der PSF ca. 25 mg beträgt, die in maximal 5 L Fermentervolumen produziert werden sollen, ist ein Detektionslimit in der Expressionskontrolle von 5 mg / L ausreichend.

Eine Zeitreihe der Expression im 2 mL Maßstab in 24-„well“ Platten wurde durchgeführt. Wie in Abb. 8A zu sehen, konnte zu allen untersuchten Zeitpunkten Protein detektiert werden. Eine Steigerung der Proteinmenge nach 24 Stunden ist im Gegensatz zum 50 mL Maßstab (siehe 4.1.2) nicht zu sehen.

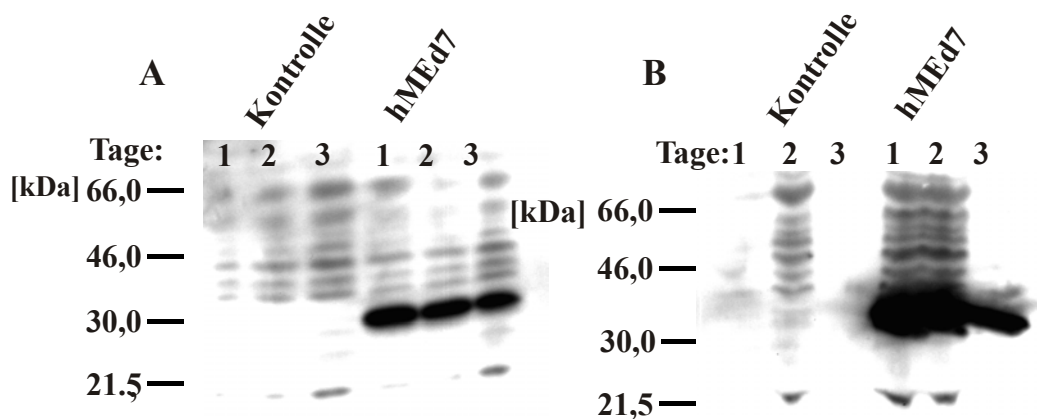


Abbildung 8: Zeitverlauf der Expression im 2 mL Maßstab in 24-„well“ Platten und Vergleich eines mechanischen Zellaufschlusses mit einer chemischen Lyse

Abb. 8A zeigt den zeitlichen Verlauf der Expression von hMed7 im 2 mL Maßstab in 24-„well“ Platten. Als Kontrolle dient ein Stamm, der mit dem leeren Vektor transformiert wurde. Die Zellen wurden mechanisch mittels Glasperlen aufgeschlossen. Die aufgetragene Menge an Zelllysat pro Spur entspricht ca. 2 OD₆₀₀. Die Proben in Abb. 8B wurden zu den gleichen Zeitpunkten gezogen wie in Teil A, die Zellen sind allerdings chemisch mittels Kochen des Pellet in SDS-PAGE Puffer aufgeschlossen. Die aufgetragene Menge pro Spur entspricht ca. 20 OD₆₀₀.

Die Zellen wurden nach dreitägiger Vorkultur in WM9-Medium mit Glukose in frisches Medium ohne C-Quelle umgesetzt und täglich 1,0 % Methanol (v/v) und 0,1 % Glukose (w/v; beides Endkonzentration) zugegeben.

Die Expression wurde mittels Western Blot unter Benutzung eines anti-Penta-His Antikörpers detektiert.

Als Größenstandard wurde „rainbow“-Marker (Amersham, Braunschweig) aufgetragen.

Um den Expressionsnachweis für eine Automatisierung zu vereinfachen, wurde eine chemische Zelllyse etabliert. Die Zellen wurden in SDS-PAGE Proben-Puffer (Laemmli 1970) gekocht. Nach einer Zentrifugation wird der Überstand direkt auf das Gel aufgetragen. Dies erleichtert eine Automatisierung, da nur Erhitzen und Zentrifugation nötig sind und das Vortexen zur mechanischen Lyse entfällt. Abbildung 8B zeigt, dass die chemische Lyse ebenfalls zu gut detektierbaren Signalen führt. Die analysierten Zellen wurden jeweils zu den gleichen Zeitpunkten geerntet wie in Abbildung 8A. Die aufgetragene Menge im Falle der chemischen Lyse entspricht einer 10fach höheren OD₆₀₀ (OD 20 im Vergleich zu OD 2). Dies ist jedoch methodisch kein Nachteil, da die geernteten Volumina sich nicht erheblich voneinander unterscheiden (750 µl bei chemischem Aufschluss zu 1 mL für mechanischen

Aufschluss). Die Unterschiede in den aufgetragenen Mengen resultieren aus den unterschiedlichen Zellkonzentrationen während des Aufschlusses.

4.2.3.5 Vergleich des Expressionslevels verschiedener cDNAs

Als Positivkontrolle sowie als Referenz zur Abschätzung der Expressionshöhe wurde der Klon 283 benutzt. Dieser stellte zu Beginn des „screening“ den Klon mit der geringsten noch detektierten Expressionshöhe dar (siehe 4.2.3.4).

Im Laufe der Arbeiten wurde festgestellt, dass zahlreiche cDNAs zu einer deutlich schwächeren Expression als der Referenzklon führten (siehe Abb. 9). Aus diesem Grunde wurde die Kategorisierung dahingehend geändert, dass der Referenzklon und Klone vergleichbarer Signalstärke als mittlere, deutlich schwächere als schlechte und deutlich bessere als gute Exprimierer eingestuft wurden. Von den 99 cDNAs, die auf Expression untersucht wurden, wurden 60 cDNAs nicht exprimiert (-), 16 schwach (+), 20 mittel (++) und drei gut (+++) exprimiert.

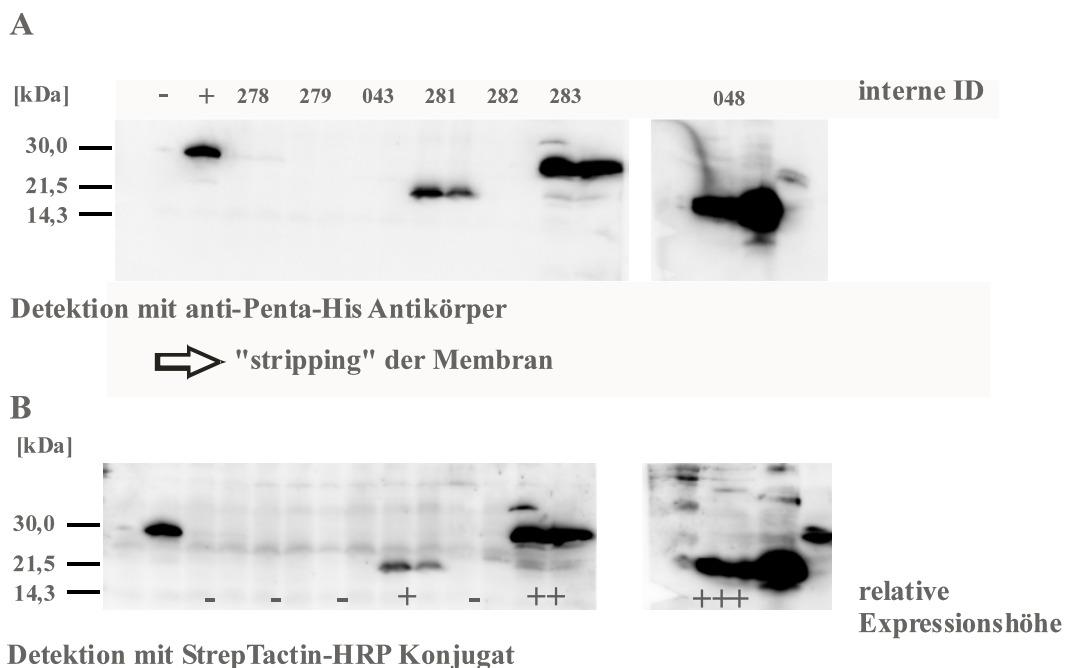


Abbildung 9: Detektion einer Auswahl der humanen Proteine mittels Western Blot

Die Klone wurden im 2 ml Maßstab drei Tage mit Glukose als C-Quelle angezogen und nach Medienwechsel 24 h induziert. Der Aufschluss erfolgte durch chemische Lyse. Nach Western Blot und Detektion mittels anti-Penta-His Antikörper wurden die Antikörper von der Membran entfernt und die Membran mit StrepTactin-HRP Konjugat inkubiert. Es wurden immer zwei Klone pro cDNA untersucht. Das „stripping“ Protokoll wurde überprüft, indem „gestrippte“ Spuren ohne neue Antikörperhybridisierung detektiert wurden.

Die aufgetragene Menge pro Spur entspricht ca. 20 OD₆₀₀. Als Größenstandard wurde „rainbow“-Marker (Amersham, Braunschweig) aufgetragen.

Durch spätere Analysen der Sequenzen stellte sich heraus, dass unter diesen Sequenzen sieben doppelt bearbeitet worden waren (siehe 4.3.2.1). Die korrigierte Anzahl beträgt damit 92 cDNAs, von denen 56 als nicht (-), 16 als schwach (+), 17 als mittel (++) und 3 als gut (+++) exprimierbar klassifiziert worden sind (siehe Tabelle 1 im Anhang).

Dies entspricht einem Anteil von 40 % exprimierbaren Proteinen.

4.2.4 Evaluierung der Affinitätstags in *P. pastoris* – Reinigung von Beispielpoteinen

Um die Funktionalität der Affinitätstags sowie den Grad der Reinigung aus *P. pastoris* Lysat und den Gehalt an rekombinantem Protein abschätzen zu können, wurden zwei Proteine jeweils über den His₆- und den StrepII-tag gereinigt. Die gereinigten Proteine waren GFP als ein sehr gut exprimiertes (siehe Abb. 11) sowie hMed7 Protein als ein Vertreter für mittlere Expression (siehe Abb.10).

Die Zellen wurden für drei Tage in WM9 / 2,0 % (w/v) Glukose angezogen. Nach dieser Vorkultur wurden die Zellen in frischem WM9 ohne C-Quelle aufgenommen und zweimal täglich mit 1 % (v/v) Methanol und 0,1 % (w/v; beides Endkonzentration im Medium) Glukose gefüttert. Nach weiteren vier Tagen wurden pro Reinigungsansatz 10 mL (bei einer OD₆₀₀ von ≈ 80) der Kultur geerntet, mittels Glassperlen aufgeschlossen und das Protein über Affinitätschromatographie gereinigt.

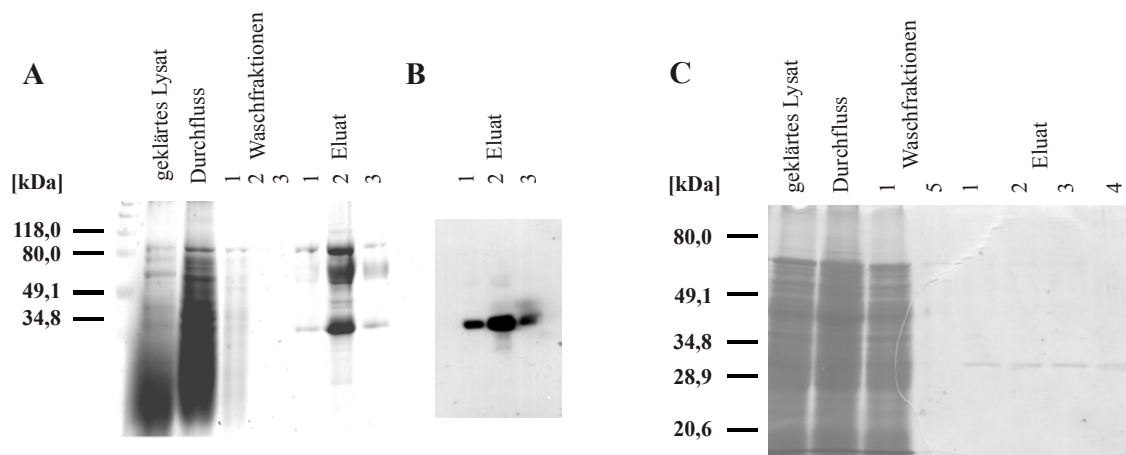


Abbildung 10: Beispiel für einstufige Reinigungen eines Proteins mittlerer Expressionsstärke

Das Protein hMed7 wurde jeweils in einem Schritt mittels Nickel-NTA- (Teil A + B) oder über StrepTactin-Matrix (Teil C) angereinigt. Jeweils 10 ml einer Kultur wurden aufgearbeitet. Bei der Ni-NTA Reinigung betragen die Eluatvolumina 200 μ l pro Fraktion, von denen jeweils 10 μ l aufgetragen wurden. Das Proteingel wurde Coomassie gefärbt (Teil A). Die Identität des Proteins wurde mittels Western Blot durch Detektion mit StrepTactin-HRP Konjugat bestätigt (Teil B). Anhand der Intensität der Bande im Coomassie gefärbten Gel wurde die Ausbeute an Protein auf 2 – 5 mg / l Kulturvolumen quantifiziert.

Teil C zeigt eine StrepTactin Affinitätschromatographie als ersten Schritt. Die Eluatvolumina bei der Reinigung über StrepTactin betragen 100 μ l. Es wurden ebenfalls 10 μ l der genannten Fraktionen auf das Gel aufgetragen. Das Gel wurde Coomassie gefärbt. Als Größenstandard wurde „broadrange prestained“-Marker (Biorad, CA) aufgetragen.

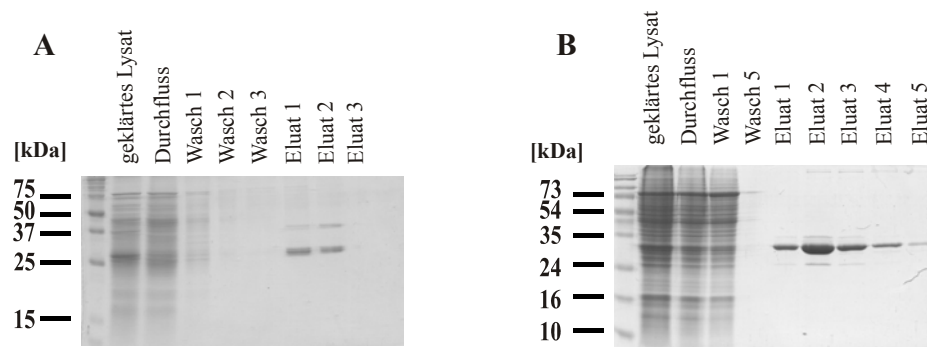


Abbildung 11: Beispiel für einstufige Reinigungen eines sehr gut exprimierten Proteins

GFP wurde einmal mittels Nickel-NTA- (Teil A) und einmal über StrepTactin-Matrix (Teil B) angereinigt. Jeweils 10 ml einer Kultur wurden aufgearbeitet. Von Lysat und Durchfluss wurden jeweils 7 μ l aufgetragen, von den Wasch- und Eluatfraktionen jeweils 10 μ l. Das Proteingel wurde Commassie gefärbt (Teil A). Es wurden von allen Fraktionen 10 μ l aufgetragen. Das Gel wurde Commassie gefärbt (Teil B). Als Größenstandard wurde in Teil A „broadrange precision“-Marker (Biorad, CA) und in Teil B „broadrange“-Marker (Biorad, CA) aufgetragen.

Beide Reinigungen zeigen, dass eine deutliche Anreicherung sowohl eines sehr gut als auch eines mittel exprimierten Proteins durch sowohl den His₆- als auch den StrepII-tag möglich ist. Die Ausbeute ist entsprechend der Expressionshöhe der Proteine unterschiedlich. Ebenso ist die Menge der verbleibenden Verunreinigungen verschieden. Im Falle des hochexprimierten GFPs ist die absolute Menge an Verunreinigungen geringer als bei hMed7. Bei der Beurteilung der Verteilung des Proteins auf die Elutionsfraktionen ist zu berücksichtigen, dass die Eluate der Nickel-NTA Reinigung das doppelte Volumen von denen der StrepTactin Reinigung haben. Zumindest im Fall des GFP zeigt die StrepTactin Reinigung eine höhere Ausbeute.

4.3 Warum werden verschiedene cDNAs derart unterschiedlich exprimiert – die Suche nach Parametern, die die Expression beeinflussen

Nach der erfolgten Expressionsanalyse von 99 cDNAs in *P. pastoris*, wurden die Sequenzen der erstellten Klonsammlung auf Eigenschaften untersucht, die Ursache für die beobachteten unterschiedlichen Expressionshöhen sein könnten. Dazu wurden im folgenden einerseits theoretische und andererseits mRNA-Analysen der Sequenzen durchgeführt. Die Sequenzdaten wurden GenBank entnommen (Benson et al. 2003, freundlicherweise zur Verfügung gestellt von Dr. C. Büsow).

4.3.1 Untersuchungen zur transkriptionellen Regulation und mRNA Stabilität

Als erster Schritt wurden Unterschiede in der mRNA-Menge oder -Stabilität untersucht. Hierzu wurde ein Teil der Expressionsklone analysiert, es wurden Vertreter aller Expressionshöhen ausgewählt.

4.3.1.1 Korrelation zwischen Expressionshöhe und der Menge der mRNA

Um vergleichbare Signale für die verschiedenen untersuchten cDNAs zu bekommen, war es nötig, diese mit der gleichen Sonde zu detektieren. Es wurde eine Sonde gegen die 5'-UTR entworfen (siehe Abb. 12). Da alle cDNAs im gleichen Vektor exprimiert wurden, beinhalten alle Transkripte die gleiche 5'UTR.

Die Sonde wurde durch den Einbau Digoxigenin markierter Nukleotide markiert. Synthetisiert wurde die Sonde mittels PCR von dem Vektor pPICHs.

Zum Vergleich der Menge an spezifischer mRNA verschiedener cDNAs, wurden Analysen einer Auswahl von Klonen mittels Northern-Blot durchgeführt.

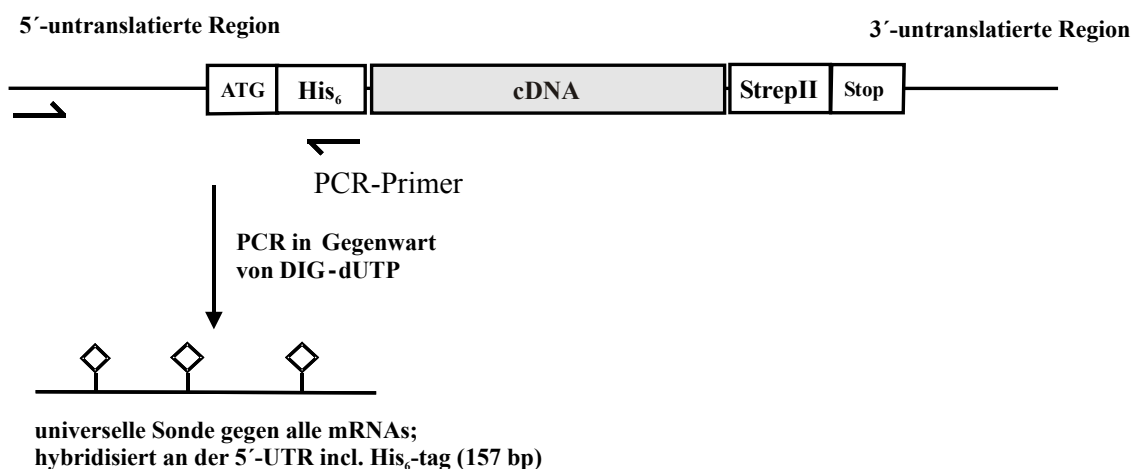


Abbildung 12: Schema der universellen Sonde zur Detektion der 5'-untranslatierte Region

Die Sonde wurde mittels PCR von der untranslatierten Region des Expressionsvektors pPICHs synthetisiert. Die Markierung erfolgte durch DIG markiertes dUTP, das während der PCR eingebaut wurde.

Abbildung 13 zeigt den Vergleich der Transkriptmengen für sechs verschiedene Expressionsklone unterschiedlicher Expressionsstärke vor und 90 min nach der Induktion. Zum einen ist zu erkennen, dass zum Zeitpunkt vor der Methanolzugabe kein Transkript detektierbar ist. Zum anderen ist 90 min nach Induktion kein Zusammenhang zwischen Expressionshöhe des Proteins und der Menge an mRNA erkennbar. So zeigt z.B. Klon 282 (-) vergleichbare Transkriptmengen wie Klon 048 (+++).

Die Tatsache, dass zumindest für diese Auswahl an Expressionsklonen kein Unterschied in der mRNA-Menge feststellbar ist, führte zu dem Ansatz, die Klone auf Unterschiede in der Stabilität der jeweiligen Transkripte zu untersuchen. Die Stabilität von Transkripten ist als ein regulatorischer Mechanismus der Genexpression in Hefe bekannt (McCarthy 1998).

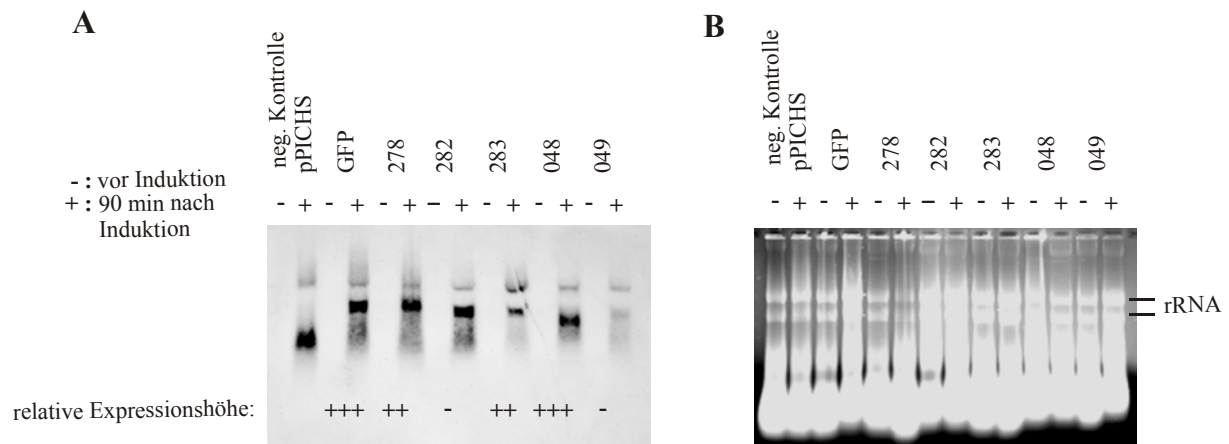


Abbildung 13: Vergleich der mRNA Mengen verschiedener cDNAs vor und 90 min nach Induktion

Es wurden Expressionsklone verschiedener Kategorien bezüglich der Transkriptmenge vor und 90 min nach Induktion verglichen. Die Klone wurden drei Tage in WM9-Medium mit 2 % (w/v) Glukose angezogen. Danach wurde durch Zugabe von 1 % (v/v) Methanol die Transkription induziert. In Teil A ist jeweils die Gesamt-RNA vor und 90 min nach Induktion aufgetragen. Die aufgetrennte RNA wurde auf Nylon Membran geblottet und die Detektion erfolgte nach Hybridisierung mit der Sonde durch Peroxidase konjugierten anti-DIG Antikörper. Teil B zeigt das zugehörige RNA-Gel vor dem Blotten. Es ist zu sehen, dass die Mengen an Gesamt-RNA pro Spur etwa gleich sind.

4.3.1.2 Untersuchungen zur Stabilität der Transkripte

Um die Halbwertszeit eines Transkriptes bestimmen zu können, muss die mRNA Synthese gestoppt werden, um dann die Abnahme des Transkriptes über die Zeit zu bestimmen. Hierzu gibt es verschiedene Möglichkeiten. In *S. cerevisiae* kann dies durch thermische Inaktivierung der Transkription bei Verwendung eines Stammes mit temperatursensitiver RNA-Polymerase II (*rpb1-1*, Herrick et al. 1990) erreicht werden. Zusätzlich kann ein reprimierbarer Promoter (*GALI*, Caponigro et al. 1993) eingesetzt werden. Eine weitere Möglichkeit besteht in der Hemmung der RNA-Polymerase II durch Thiolutin (Das et al. 2000).

Da eine temperatursensitive *P. pastoris* RNA-Polymerase II Mutante nicht zur Verfügung steht und der *AOX1*-Promoter als Glukose reprimiert beschrieben ist (Tschopp et al. 1987) wurde zunächst untersucht, ob diese Repression ausreicht, auch in Gegenwart von Methanol die Transkription zu unterbinden. In diesem Fall wäre es möglich, die Transkription in induzierten Kulturen mittels Zugabe von Glukose zu stoppen.

4.3.1.2.1 Untersuchungen zur Repression des *AOX1*-Promotors durch Glukose

Anhand des Expressionsklons 048, eines Expressionsklons für GFP sowie eines Kontrollstammes mit dem Vektor pPICHS wurde die Repression des Promotors durch

Glukose untersucht. Alle Stämme zeigten detektierbare Mengen an Transkript (siehe Abb. 13).

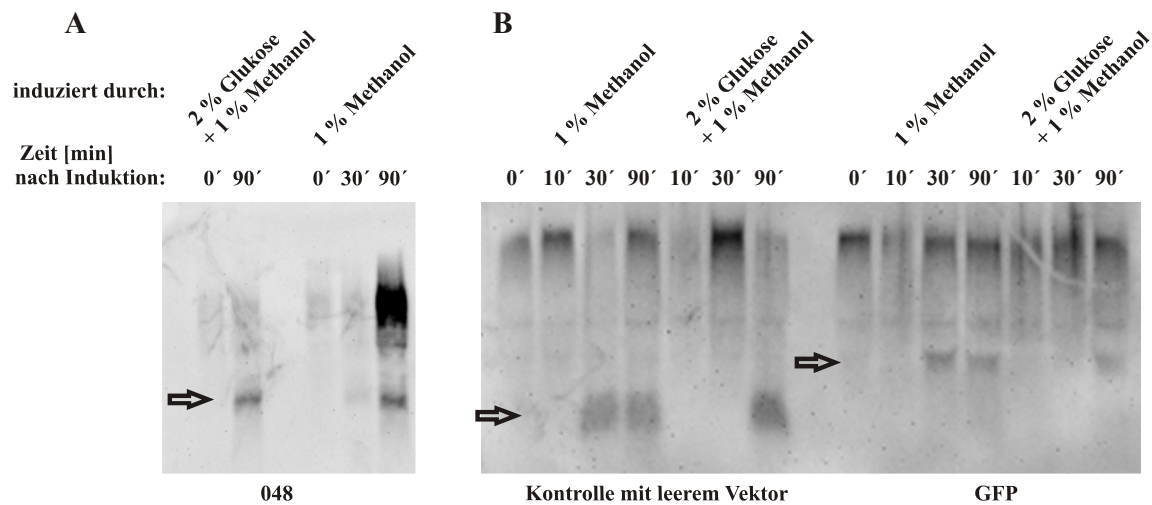


Abbildung 14: Die Repression des *AOX1*-Promotors durch Glukose in Gegenwart von Methanol

Die drei Stämme wurden drei Tage in WM9 Medium mit 2,0 % (w/v) Glukose angezogen. Die Kulturen wurde in zwei Hälften geteilt. In einem Teil wurde die Transkription durch Zugabe von 1,0 % (v/v) Methanol induziert, dem anderen Teil 1,0 % (v/v) Methanol und 2,0 % (w/v) Glukose zugesetzt. Zu den genannten Zeitpunkten wurden Proben gezogen und das Transkript (⇒) wurde per Northern-Blot detektiert. Die erwarteten Größen der Transkripte sind 809 Bp (048), 387 Bp (Kontrolle mit leerem Vektor) und 1092 Bp (GFP). Die gleichmäßige Beladung der Spuren wurde anhand des Gels kontrolliert (nicht gezeigt).

Für alle drei Stämme gilt, dass die cDNA spezifische Transkriptmenge nach 90 Minuten in beiden Ansätzen gleich ist (siehe Abb. 14). Hieraus folgt, dass eine Repression des Plasmidpromotors durch die eingesetzte Glukosemenge in Gegenwart von Methanol nicht vorliegt. In den Experimenten mit dem GFP exprimierenden Klon sowie dem Kontrollklon mit dem leeren Vektor wurde der Zeitverlauf der Transkript Akkumulation feiner aufgelöst (0', 10', 30' und 90'). Wie aus Abb. 14B ersichtlich ist, ist für beide Klone in den 30 min-Proben nach Zugabe von Glukose und Methanol kein Transkript zu sehen, während in den Kulturen, die nur mit Methanol versetzt wurden, noch Transkript vorhanden ist. Weiterhin ist in den Kulturen, denen nur Methanol zugesetzt worden ist, keine Veränderung in der Transkriptmenge zwischen 30 min und 90 min zu sehen. Dies deutet darauf hin, dass durch entsprechende Mengen Glukose im Medium entweder der Start der Transkription verzögert wird oder die Transkriptionsrate verlangsamt wird, so dass die Gleichgewichtskonzentration an mRNA später erreicht wird. Dass nach 90 min die Glukose im Medium verbraucht ist, ist unwahrscheinlich, da die „crabtree“-negativen Stämme auch nach zwei Tagen mit der gleichen Menge noch deutliches Wachstum zeigen.

4.3.1.2.2 Die Repression des *AOX1*-Promotors durch Glukose in Abwesenheit des Induktors

Da der *AOX1*-Promoter auf dem Plasmid durch Glukose in methanolhaltigem Medium nicht ausreichend reprimiert wird, musste die Strategie modifiziert werden. Da in Medium ohne Methanol keine Transkription stattfindet (siehe Abb. 13), wurden die Kulturen 90 min nach der Induktion geteilt, geerntet und die Zellen in frischem Medium ohne C-Quelle resuspendiert. Unmittelbar danach werden 1,0 % (v/v) Methanol bzw. 2,0 % (w/v) Glukose zugesetzt. Der Verlauf der Transkriptmenge wurde verfolgt (siehe Abb. 15).

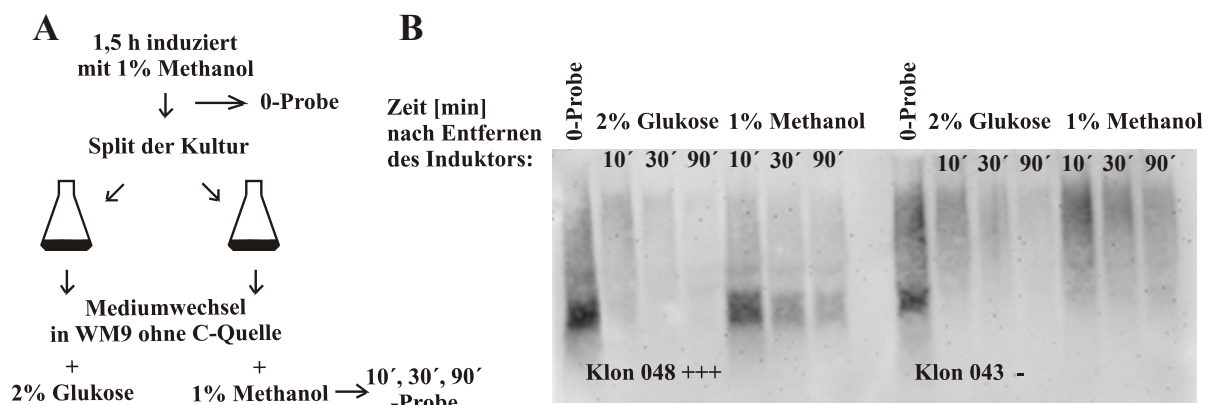


Abbildung 15: Stabilität der Transkripte nach Wechsel in frisches Medium

Untersucht wurde die Transkriptabnahme nach Entfernung des Induktors durch Wechsel des Mediums. Teil A zeigt die Durchführung des Versuchs. Zu den genannten Zeitpunkten wurden Proben gezogen, Gesamt-RNA isoliert und mittels Northern-Blot das Transkript detektiert (Teil B). Die untersuchten Beispiellone waren 048 als guter Exprimierer und 043 als Nichtexprimierer. Die erwarteten Größen der Transkripte sind 809 Bp (048) und 899 Bp (043). Die gleichmäßige Beladung der Spuren wurde anhand des Gels kontrolliert (nicht gezeigt).

In Abbildung 15 ist zu sehen, dass in beiden Stämmen zum Zeitpunkt der 0-Probe (90 min nach Zugabe von Methanol, vor dem Split der Kulturen und dem Ernten der Zellen) die gleiche Menge an Transkript vorliegt. In beiden Stämmen ist bereits 10 min nach Resuspension in glukosehaltigem Medium kein Transkript mehr detektierbar. Dies zeigt eine Halbwertszeit im Bereich von weniger als zehn Minuten. Weiterhin ist zu erkennen, dass bei Klon 048 das Transkript in frischem methanolhaltigem Medium über die beobachteten 90 min stetig abnimmt. Dies ist unerwartet, da der Transkriptgehalt der 90 min Proben, dem der 0-Probe vor dem Medienwechsel entsprechen sollte. Dies deutet auf eine verminderte Syntheserate in frischem Medium hin. Unter der Annahme, dass die Halbwertszeit in frischem Medium der in verbrauchtem entspricht, würde dann die Konzentration abnehmen. Interessanterweise ist bei Klon 043 (dem Nichtexprimierer) bereits 10 min nach dem Wechsel auf frisches, methanolhaltiges Medium kein Transkript mehr zu sehen.

Möglicherweise liegt das an unterschiedlichen Halbwertszeiten der beiden untersuchten Transkripte.

Da diese Beobachtungen unter induzierenden Bedingungen stattfanden und somit eine Gleichgewichtskonzentration beobachtet wird, ist die Syntheserate zu berücksichtigen. Nur unter der Voraussetzung, dass diese für beide Transkripte gleich ist, kann auf die Halbwertszeit geschlossen werden. Unter diesen Bedingungen kann dies auch nur qualitativ geschehen.

4.3.1.2.3 Austausch von Glutamat gegen NH_4Cl im Medium

Aufgrund der Abnahme des Transkriptes in frischem Medium unter induzierenden Bedingungen (siehe 4.3.1.2.2), wurde das Medium hinsichtlich weiterer Komponenten untersucht, die reprimierend wirken könnten. Als mögliche reprimierende C-Quelle kommt hier das Kohlenstoffgerüst von Glutamat in Frage. Glutamat wurde unter Beibehaltung der Molarität von Stickstoff gegen NH_4Cl ausgetauscht. Mit diesem Medium wurde das Experiment wie unter 4.3.1.2.2 beschrieben, wiederholt. Sowohl die Vorkulturen als auch die Experimente zum Abbau der mRNA erfolgten auf glutamatfreiem Medium. Aufgrund der Erkenntnisse über die kurze Halbwertszeit der Transkripte wurde eine zusätzliche Probe fünf Minuten nach dem Medienwechsel gezogen.

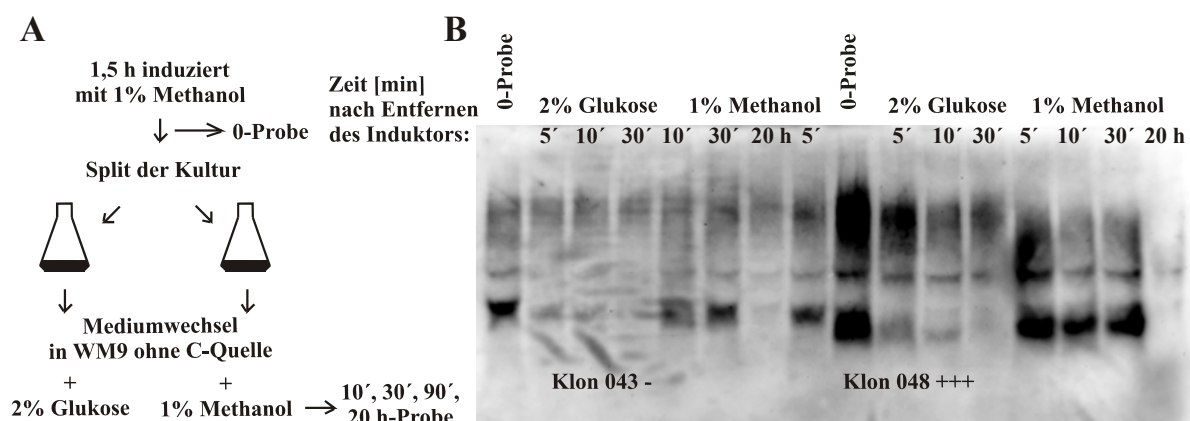


Abbildung 16: Stabilität der Transkripte in Glutamat-freiem Medium

Untersucht wurde, ob die Transkriptabnahme nach Entfernung des Induktors durch Wechsel des Mediums vergleichbar zu der in Glutamathaltigem Medium verläuft. Teil A zeigt die Durchführung des Versuchs. Zu den genannten Zeitpunkten wurden Proben gezogen, Gesamt-RNA isoliert und mittels Northern-Blot das Transkript detektiert (Teil B). Die untersuchten Beispielklone waren 048 als guter Exprimierer und 043 als Nichtexprimierer. Alle Anzuchten erfolgten in Medium mit NH_4Cl anstelle von Glutamat. Die erwarteten Größen der Transkripte waren 899 Bp (043) und 809 Bp (048). Die gleichmäßige Beladung der Spuren wurde anhand des Gels kontrolliert (nicht gezeigt).

Aus Abbildung 16 ist zu ersehen, dass für beide Stämme in Medium ohne Glutamat die Menge an Transkript bis zum Zeitpunkt 30 min nach Induktion stabil bleibt. Glutamat – oder ein Abbauprodukt - scheint also den *AOX1*-Promoter zu reprimieren. Eine 90 min Probe

wurde in diesem Fall nicht analysiert. 20 h nach Medienwechsel ist in methanolhaltigem Medium kein Transkript mehr zu detektieren.

In den Proben, die fünf Minuten nach Wechsel auf glukosehaltiges Medium genommen wurden, ist für beide Stämme ein schwaches Signal zu sehen. Dieses Signal ist im Vergleich zur 0-Probe zu schwach, um eine Halbwertszeit zu bestimmen. Es kann kein Unterschied in der Halbwertszeit zwischen den beiden untersuchten Stämmen festgestellt werden. Klon 048 zeigt ein in allen positiven Proben stärkeres Signal. Da die Menge des jeweils gezogenen Kulturvolumens von einer OD₆₀₀-Messung abhing, die unmittelbar vor dem Medienwechsel durchgeführt wurde (abgesehen von der 20 h Probe) würde eine Ungenauigkeit dieser Messung sich durch alle entsprechenden Proben ziehen.

Der Zeitpunkt der ersten Probenahme nach fünf Minuten stellt eine untere Grenze für die analysierbare Zeitspanne dar. Aufgrund der Tatsache, dass die Zellen in frischem Medium zunächst gleichmäßig resuspendiert werden müssen, um eine vergleichbare Probennahme zu gewährleisten, ist diese Zeitspanne technisch nicht wesentlich zu verringern. Eine Messung von Halbwertszeiten ist also in dem sich hier abzeichnenden Zeitbereich mit der dargestellten Methode nicht durchführbar. Allgemein kann jedoch gesagt werden, dass die Transkripte von den analysierten Konstrukten instabil sind. Selbst ein Transkript, das zu guten Expressionshöhen auf Proteinebene führt (048), hat eine Halbwertszeit von unter fünf Minuten.

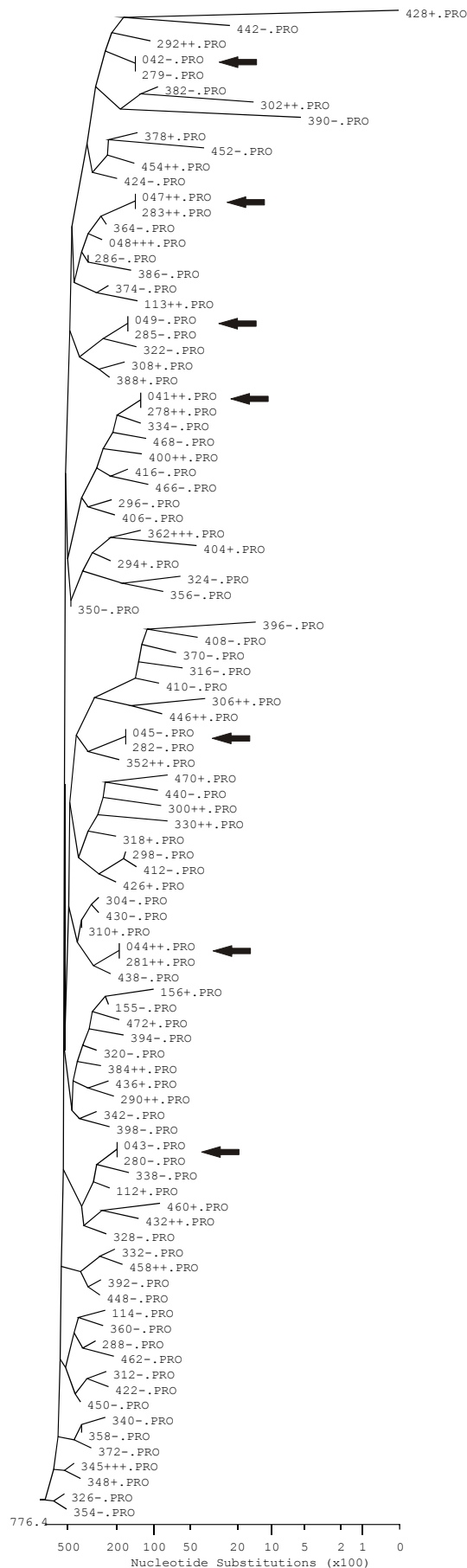
Weitere mRNA Analysen, z.B. Halbwertszeitbestimmungen mittels Hemmung des Promotors durch Thiolutin oder Ausweitung der Analysen auf eine größere Anzahl von Expressionsklonen, konnten innerhalb dieser Arbeit nicht mehr durchgeführt werden.

4.3.2 Ungerichtete Suche nach Sequenzmerkmalen, die mit der Expressionshöhe korrelieren

Im folgenden wurden die Sequenzen mittels bioinformatischer Anwendungen auf Merkmale hin untersucht, die mit einer der Kategorien der Expressionshöhe korrelieren. Ziel war es, sequenzbasierte Parameter zu identifizieren, die mit einer der Kategorien der Expressionshöhe korrelieren, oder eine Vorhersage zur Expression einer gegebenen Sequenz erlauben.

Zunächst wurden die Sequenzen ungerichtet, d.h. nicht mit Blick auf ein bestimmtes Merkmal analysiert. Hierzu wurden zum einen „multiple Alignments“ aller Sequenzen und weiterhin phylogenetische Bäume erstellt, zum anderen die Sequenzen auf gemeinsame Motive hin untersucht.

4.3.2.1 Gemeinsamkeiten über phylogenetische Bäume und Sequenzabhängigkeit der Expressionshöhe



Um zu sehen, ob die Expressionshöhe von globalen Gemeinsamkeiten der Sequenzen abhängt, wurden phylogenetische Bäume erstellt. Hierzu wird zunächst ein multiples Alignment aller Sequenzen erstellt und diese werden dann aufgrund ihrer Ähnlichkeiten im Baum angeordnet. Dies wurde sowohl auf DNA als auch auf Proteinebene durchgeführt. Es ist zu bedenken, dass die analysierten Sequenzen nicht verwandt sind und Alignments bei den hier vorhandenen geringen Ähnlichkeiten als fehlerträchtig angesehen werden müssen (Mount 2001). Trotzdem wurde versucht, ob diese Methode ausreicht, Gemeinsamkeiten innerhalb der Kategorien zu identifizieren. Es wurden verschiedene Parametereinstellung zur Erstellung der multiplen Alignments verwandt. Ein Beispiel für einen Baum ist in Abb. 17 gezeigt. Es ist zu sehen, dass keine Cluster – hier: gemeinsame Äste – von Sequenzen mit gleicher Expressionshöhe (die Einstufung ist aus den + bzw. – im Sequenznamen zu erkennen) auftreten.

Abbildung 17: Beispiel eines phylogenetischen Baumes

Dargestellt ist ein phylogenetischer Baum der Proteinsequenzen. Das zugrundeliegende Alignment wurde mittels des Algorithmus ClustalW erstellt. Die eingestellten Parameter waren: „gap penalty“ 10, „gap length penalty“ 10. Die Substitutionsmatrix war PAM250. Die identischen Sequenzen in der Analyse sind durch Pfeile markiert.

Es waren mehrfach auftretende identische cDNAs in der Analyse, die bei der Vorauswahl (siehe 4.2.3.1) nicht aussortiert worden sind und, da sie aus verschiedenen Eingangsklonen stammten, in der Folge als verschiedene cDNAs behandelt worden und teilweise auch in verschiedenen Chargen von Anzucht und Aufarbeitung bearbeitet worden sind. Im Alignment wurden diese als gleich erkannt und im Baum an den Enden der Zweige angeordnet (z.B. 042-.pro und 279-.pro oder 047++.pro und 283++.pro; siehe Pfeile in Abb. 17).

Alle gleichen Sequenzen wurden im Expressionstest in die gleichen Kategorie der Expressionshöhe einsortiert. Dies zeigt die gute Reproduzierbarkeit des Expressionstests und der Einstufungen. Weiterhin bestätigt dies die Sequenzabhängigkeit der Expressionshöhe.

4.3.2.2 Suche und Charakterisierung von Sequenzmotiven

Für die nachfolgenden Analysen wurde der Satz der Sequenzen um die doppelt vorhandenen auf 92 reduziert. Von diesen wurden 56 als nicht (-), 16 als schwach (+), 17 als mittel (++) und 3 als gut (+++) exprimierbar klassifiziert (siehe Tabelle 1 im Anhang).

Um das unter 4.3.2.1 erwähnte Problem der Vergleichbarkeit sehr verschiedener Sequenzen zu lösen, wurde im folgenden nach Sequenzmotiven gesucht. Zu diesem Zwecke wurde das Programm MEME genutzt. MEME findet gemeinsame Motive in einem Satz von Sequenzen auch, wenn nur ein Teil dieser Sequenzen dieses Motiv enthält (Bailey 1994). Die Analyse wurde mit allen Proteinsequenzen (ohne Tags) durchgeführt. Aufgrund der Analyse eines Kontrolldatensatzes aus zufälligen Sequenzen mit gleicher Länge und Zusammensetzung wie der zu analysierende Datensatz wurde der Grenzwert für den „P-value“ auf $1 \cdot 10^{-30}$ festgelegt. Es wurden 18 Motive gefunden, von denen mindestens ein Vertreter einen kleineren „P-value“ aufwies. Zunächst wurde das Ergebnis auf Motive untersucht, die lediglich in einer Kategorie der Expressionshöhe vorkommen. Es wurden 11 Motive gefunden (siehe Tabelle 2 im Anhang), die alle in nichtexprimierten Sequenzen vorkamen. Die Motive verteilen sich nicht gleichmäßig auf die Proteine, sondern liegen in Gruppen vor. Die Motive 2, 5, 7, 8, 9, 10 und 11 kommen immer in Kombination und nur in zwei Proteinen vor, die Motive 1, 3, und 4 kommen in Kombination in drei Proteinen vor, ein weiteres Protein enthält nur Motiv 4. Um zu sehen, ob die gefundenen Motive als biologisch bedeutsam charakterisiert sind, wurde in PROSITE nach diesen Motiven gesucht. Bereits die Eingabe von Motiv 1 ergab einen Eintrag (PROSITE Acc. No. PS00421): das Motiv ist beschrieben als intrazellulärer „loop“ von Proteinen mit vier Transmembrandomänen. Die Kombination der Motive 1, 3 und 4 kennzeichnet Proteine der Superfamilie mit vier Transmembrandomänen (TM4SF), das Protein mit dem einzelnen Motiv 4 ist ebenfalls ein Membranprotein. Das Motiv hat keinen

Eintrag in PROSITE, scheint aber mit Membranproteinen assoziiert zu sein. Die zweite Motivkombination (2, 5, 7, 8, 9, 10 und 11) gehört zu zwei Adenylyl-Zyklase assoziierten Proteinen. Hier kann nicht gefolgert werden, dass die gefundenen Motive die Expressionshöhe bedingen. Die ähnliche Funktion der Proteine kann ebenso gut Ursache für die Gemeinsamkeiten sein.

Entsprechend der Vorgaben des Teilprojektes Bioinformatik der PSF sollten keine Membranproteine in der Auswahl enthalten sein (siehe 4.2.3.1). Da die Expression von Membranproteinen zusätzlichen Limitationen unterliegt, die z.B. durch den intrazellulären Transport bedingt sind (Butz et al. 2003), sollte der Fokus dieser Arbeit auf Expression von löslichen, intrazellulären Proteinen liegen (siehe 2.2). Da die gefundenen Motive teilweise als charakteristisch für Membranproteine bekannt waren, wurden zunächst alle in dieser Analyse befindlichen Zielproteine mittels SwissProt auf annotierte Lokalisation überprüft. Für 12 der 56 nicht exprimierten Proteine ergab sich eine Annotation als integrales Membranprotein, resultierend entweder aus in der Sequenz ersichtlichen Transmembranomänen oder aus experimentellen Daten. Zwei weitere sind aufgrund von Sequenzähnlichkeiten als Zellmembran-lokalisiert annotiert. In der Kategorie der schwachen Exprimierer ist ein Protein als integrales Membranprotein, ein weiteres über Sequenzähnlichkeit als Membranprotein annotiert. Von den mittel exprimierten Proteinen sind drei Proteine als integrale Membranproteine annotiert, während unter den gut exprimierten Sequenzen kein Membranprotein zu finden ist (siehe Tabelle 1 im Anhang).

In Zusammenarbeit mit dem Teilprojekt Bioinformatik der PSF wurden in den in SwissProt als integrale Membranproteine annotierten Sequenzen Transmembranomänen in der Sequenz gefunden. In den in SwissProt über Sequenzähnlichkeiten annotierten Proteinen sind *in silico* keine Transmembranomänen erkennbar (Dr. B. Simon, persönliche Mitteilung). Im folgenden wurden die Proteine mit Transmembranomänen aus der Analyse herausgenommen. Nach Reduzierung des Datensatzes um die eindeutigen Membranproteine ergibt sich eine Verteilung von 44 nicht exprimierten Proteinen, 15 schlecht, 14 mittel und drei gut exprimierten Proteinen.

Mit dem korrigierten Datensatz wurde eine erneute MEME-Analyse durchgeführt. Das Ergebnis wurde auf Motive, die mit der Expressionshöhe, sowie auf Motive, die mit genereller Exprimierbarkeit assoziiert sind, untersucht. Es ergaben sich keine neuen Motive, die einen besseren „P-value“ als die Kontrolle aufwiesen.

4.3.3 Gerichtete Analyse sequenzbasierter Parameter und deren Verteilung auf die Kategorien der Expressionshöhe

Da im Vergleich der Sequenzen miteinander keine allgemeinen Sequenzmerkmale bzw. Motive gefunden werden konnten, die mit der Expressionshöhe korrelieren, wurden die einzelnen Sequenzen auf Nukleotid- und Proteinebene im folgenden auf bekannte Sequenzmerkmale wie z.B. „codon usage“ hin untersucht und deren Verteilung auf die Kategorien der Expressionshöhe analysiert. Die statistische Signifikanz einer Korrelation dieser Merkmale mit den Kategorien der Expressionshöhe wurde auf dem 5 %-Niveau mittels des Kruskal-Wallis-Test überprüft. Die Parameterverteilungen sind im folgenden als Boxplots dargestellt. Neben den einzelnen Datenpunkten – jeder Punkt steht für eine Sequenz – sind der Median sowie oberes und unteres Quartil dargestellt.

Um eine aussagekräftigere Verteilung der Klone auf die Kategorien zu erhalten, also mehr Klone in der Kategorie der guten Exprimierer zu erhalten, wurden gezielt cDNAs, die in *S. cerevisiae* gut exprimiert wurden, für die Expression in *P. pastoris* umkloniert. Fünf Sequenzen wurden ausgesucht (Dr. C. Holz, persönliche Mitteilung). Von diesen führten drei zu einer guten Expression in *P. pastoris* (siehe Tabelle 1 im Anhang) und wurden im folgenden in die Analyse einbezogen.

4.3.3.1 AT-reiche Cluster und GC-Gehalt der Sequenzen

Wie in Abbildung 18A zu sehen, ist eine gute Expressionshöhe mit einem geringen Gehalt an AT-reichen Regionen assoziiert. Die anderen drei Kategorien unterscheiden sich untereinander nicht signifikant, enthalten jedoch alle einen höheren Anteil an AT-reichen Sequenzen im Vergleich zur Gruppe der gutexprimierten cDNAs. Die Assoziation bleibt erhalten, wenn man das Maß für den Gehalt an AT-reichen Bereichen auf die Länge der Sequenz bezieht (Abb. 18B). Dieser Trend kann weder im GC-Gehalt der gesamten Sequenzen (Abb. 18C) noch im GC-Gehalt an dritten synonymen Positionen (GC3s; Abb. 18D) beobachtet werden. Die Korrelation von guter Expression mit niedrigem Gehalt an AT-reichen Bereichen ist also kein Artefakt bedingt durch höheren AT-Gehalt oder Besonderheiten der „codon usage“.

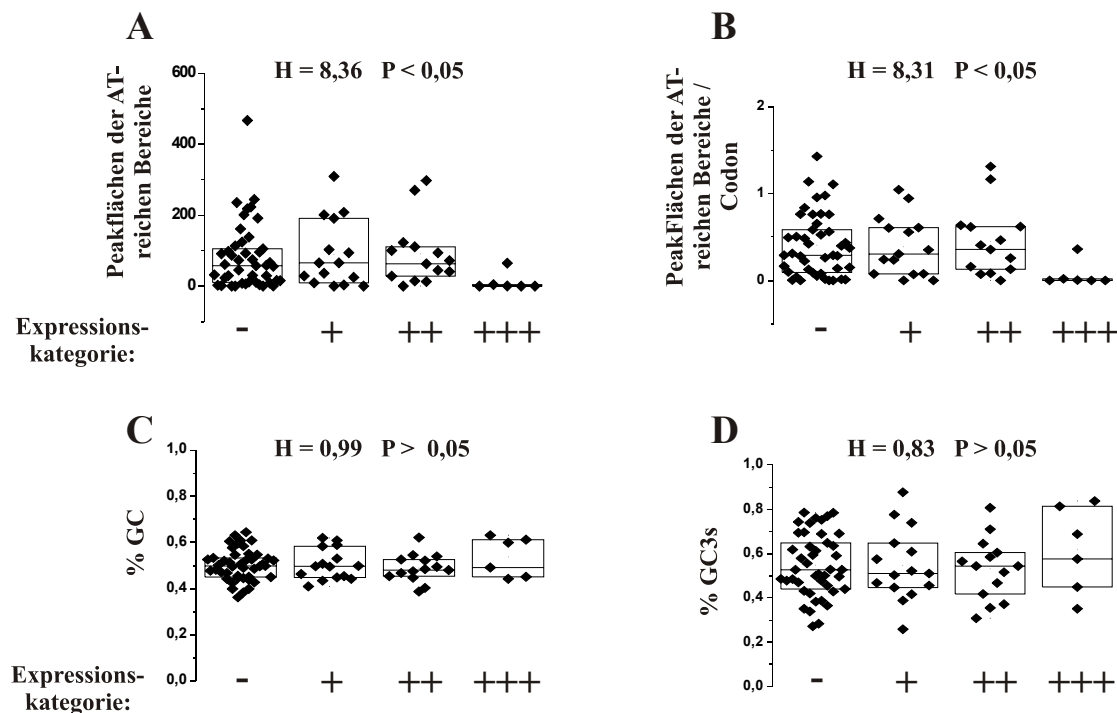


Abbildung 18: Verteilung von AT-Clustern, GC-Gehalt und GC3s auf die Expressionshöhen

Die Peakfläche der AT-reichen Bereiche korreliert auf dem 5 %-Niveau mit einer guten Expression ($P < 0,05$). Diese Korrelation ist ebenfalls vorhanden, wenn die Peakflächen auf die Länge der Sequenzen bezogen werden. Für den GC-Gehalt und den GC-Gehalt an der dritten synonymen Codonposition ist keine Korrelation feststellbar ($P > 0,05$).

4.3.3.2 Untersuchungen zur Verteilung der „codon usage“

Um die „codon usage“ zu messen, wurden der „codon adaptation index“ (CAI) und die Anzahl der effektiven Codone (Nc) berechnet. Diese Indices wurden gewählt, weil sie – im Gegensatz zu anderen existierenden Indices – vergleichbare Werte unabhängig von der Länge der analysierten Sequenz liefern (Comeron und Aguade 1998). Der Nc kann Werte zwischen 20,0 (maximal möglicher Trend in der „codon usage“) und 61,0 (keine Bevorzugung bestimmter Codone) annehmen. Der CAI wurde basierend auf einem Set von 24 hochexprimierten Genen aus *S. cerevisiae* (Sharp und Cowe 1991) kalkuliert. Der Index läuft von 0 (keine Anpassung an den Referenzorganismus) bis 1 (maximale Anpassung). Neuere Untersuchungen, die auf den deutlich zugenommenen genomischen Informationen basieren, zeigen, dass durch Einbeziehung der aktuellen Daten keine signifikante Veränderung der resultierenden Werte erfolgt und somit die Zusammenstellung von Sharp und Cowe eine gute Repräsentation darstellt (Jansen et al. 2003).

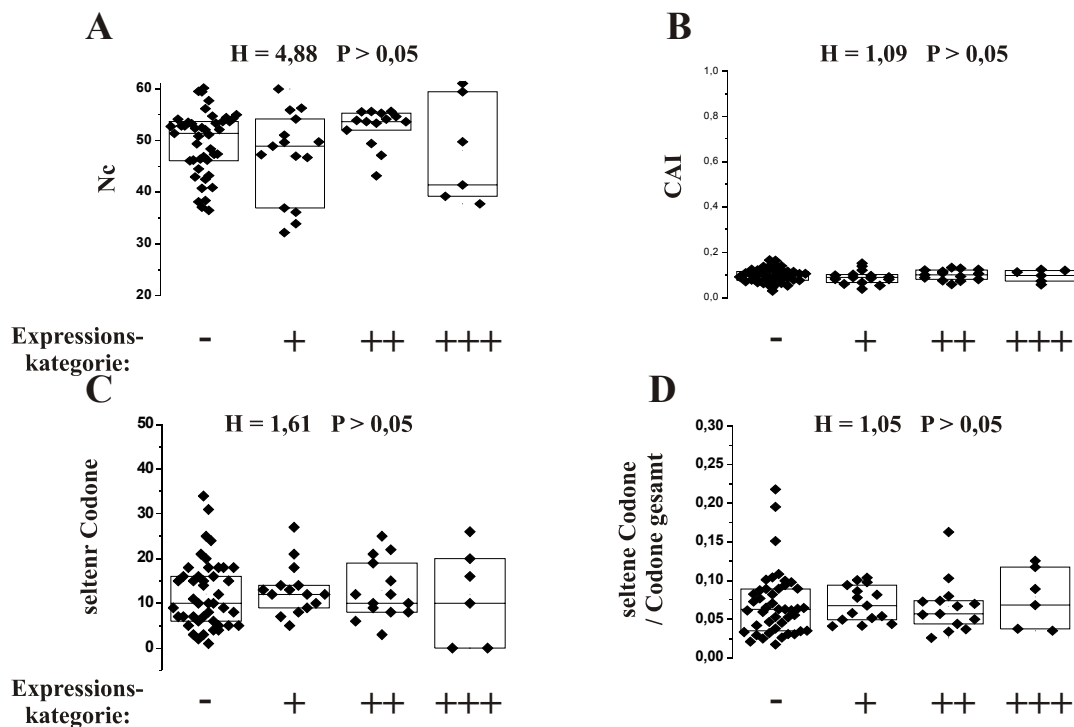


Abbildung 19: Verteilung von Parametern der „codon usage“ und von seltenen Codonen

Teil A zeigt die Verteilung der Anzahl effektiver Codone (Nc), ein allgemeines Maß für Trends in der „codon usage“. In Teil B ist die „codon adaptation“ in Bezug auf *S. cerevisiae* abgebildet. Teil C und D zeigen die Verteilung seltener Codone, absolut (C) und bezogen auf die Genlänge (D). „Selten“ bezieht sich auf die „codon usage“ in *S. cerevisiae*.

Es wurde gezeigt, dass, zumindest für die bekannten *P. pastoris* Gene, die hoch exprimiert werden, die bevorzugten Codone denen von *S. cerevisiae* ähneln (Sinclair und Choy 2002). Eine Messung des CAI gegen ein Set von *S. cerevisiae* Genen ist deshalb ein gültiger Ansatz. Abbildung 19 zeigt, dass eine Korrelation der Expressionshöhe weder mit der Codon Verteilung (Nc, Abb. 19A) noch mit der „codon adaptation“ (CAI, Abb. 19B) existiert. Teil C und D zeigen die Verteilung der Anzahl von seltenen Codonen. Es ist weder ein Zusammenhang zwischen einer der Kategorien der Expressionshöhe mit der Anzahl der seltenen Codone in der cDNA, noch mit der Frequenz von seltenen Codonen zu sehen. Es ist kein Einfluss der „codon usage“ auf die Expressionshöhe feststellbar.

4.3.3.3 Allgemeine Proteineigenschaften

Die berechneten Proteineigenschaften sind Durchschnittswerte für jedes Protein über die gesamte Sequenz. Von diesen Werten korreliert weder Hydrophobizität (Abb. 20A) noch Aromatizität (Abb. 20B) noch Protein Länge (Abb. 20C) mit der Expressionshöhe. Eine signifikante Korrelation existiert aber zwischen der Nichtexprimierbarkeit und einem hohen isoelektrischen Punkt (pI; Abb. 20D).

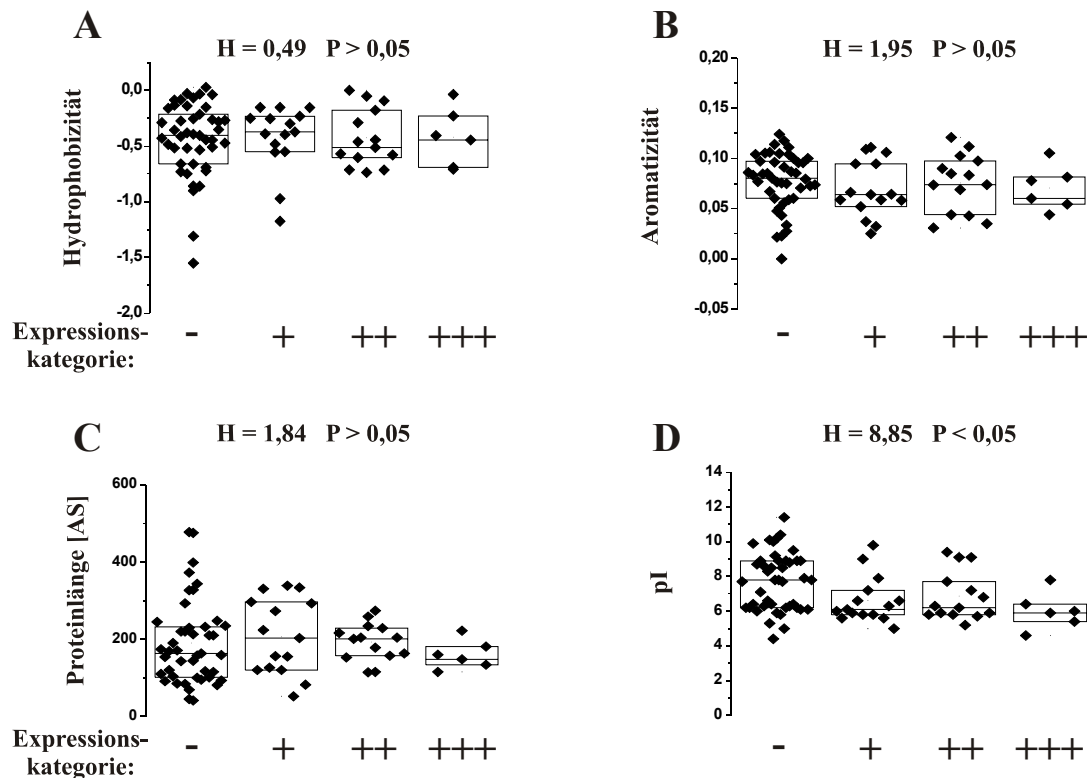


Abbildung 20: Verteilung allgemeiner Proteineigenschaften

Dargestellt ist die Verteilung allgemeiner Proteineigenschaften in den Kategorien der Expressionshöhen. Die Werte sind Durchschnittswerte über die gesamte Proteinsequenz. Ein hoher isoelektrische Punkt (D) korreliert mit keiner Expression des Proteins.

Der Median der pIs in der Kategorie der nicht exprimierten Proteine liegt bei 7,8 im Gegensatz zu den Medianen der anderen Kategorien, die zwischen 6,0 und 6,3 liegen. Ein hoher pI scheint sich somit negativ auf eine Expression in *P. pastoris* auszuwirken.

4.3.3.4 Proteindegradationssignale

Um mögliche Unterschiede in der Halbwertszeit der Proteine *in vivo* zu betrachten, wurden Sequenzparameter, die die Proteininstabilität in Hefe beeinflussen, betrachtet.

4.3.3.4.1 PEST-Motive

Das PEST-Motiv ist ein gut charakterisiertes Sequenzmotiv, das in Proteinen mit geringen Halbwertszeiten *in vivo* zu finden ist (Rogers et al. 1986). Dieses hydrophile Motiv kann Proteine in diversen Spezies destabilisieren. Es gibt Hinweise, dass das PEST-Motiv Proteine einer Degradation durch das 26S-Proteasom zuführt (zusammengefasst von Rechsteiner und Rogers 1996). Das benutzte Programm PESTfind (Rechsteiner und Rogers 1996) findet PEST ähnliche Sequenzen und weist diesen einen Wert ansteigend mit der Ähnlichkeit zum Konsensus zu.

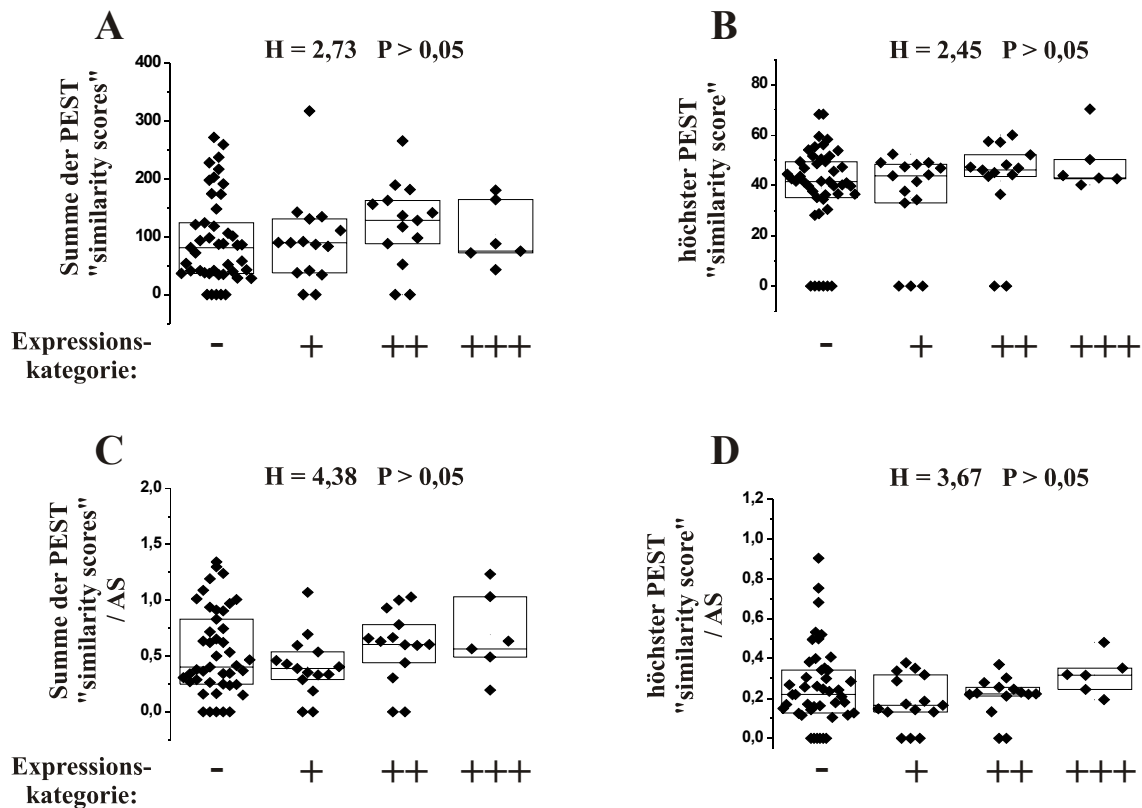


Abbildung 21: Verteilung von PEST-Motiven

Dargestellt ist die Verteilung von PEST-Motiven in den Kategorien der Expressionshöhen. Teil A zeigt die Summe der „similarity scores“ – also die Ähnlichkeit zum Consensus – von allen pro Sequenz gefundenen PEST ähnlichen Motiven. Teil B zeigt jeweils nur den höchsten „similarity score“ der je Protein gefunden wurde. In Teil C und D sind diese Werte auf die Längen der Proteine bezogen. Proteinen, in denen kein PEST ähnliches Motiv gefunden wurde, wurde der Wert Null zugeordnet.

Nach einer Transformation der Ausgabe-Werte in den positiven Bereich ist ein Wert von 50 und größer per Definition ein putatives PEST-Motiv und Motive mit Werten größer 55 sind wahrscheinlich wirksame Motive (Rechsteiner und Rogers 1996). Aufgrund der Tatsache, dass zum Teil mehrere PEST ähnliche Motive pro Protein auftreten, gibt es Sequenzen mit mehr als einem Wert. Derzeit ist nicht bekannt, wie die Wechselwirkung mehrerer PEST Motive in einem Protein ist. Aus diesem Grunde wurde die Summe aller gefundenen Motive betrachtet (Abb. 21A) sowie jeweils nur der höchste Wert je Protein – also das Motiv, das dem Consensus am ähnlichsten ist (Abb. 21B). Beide Werte wurden ebenfalls bezogen auf die Proteinlänge betrachtet (Abb. 21C und D). Eine Korrelation zwischen einem der betrachteten Parameter zu einer der Expressionskategorien konnte nicht festgestellt werden. Wenn Proteindegradation *in vivo* zu reduzierten Proteinausbeuten führt, kann diese nicht hauptsächlich auf PEST-Motiv gesteuerte Degradation zurückgeführt werden.

4.3.3.4.2 Lysinreste als Stellen für eine Ubiquitinierung

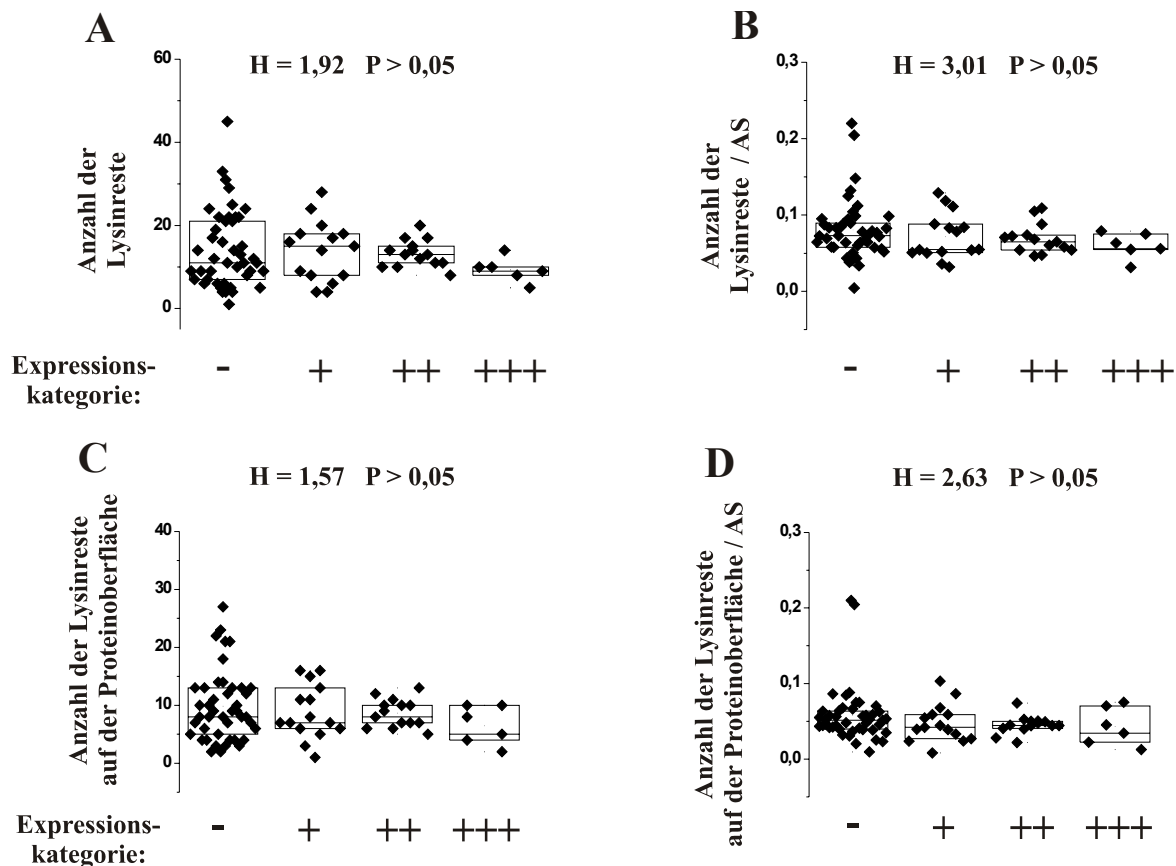


Abbildung 22: Verteilung von Lysinresten

Dargestellt ist die Verteilung der absoluten Anzahl an Lysinresten im Protein (A) und die Anzahl von Lysinresten bezogen auf die Proteinelänge (B). In Teil C und D sind nur Lysinreste berücksichtigt, die wahrscheinlich an der Proteinoberfläche liegen (Emini Index ≥ 1). Die Parameter korrelieren nicht mit der Expressionshöhe.

Eine alternative Herangehensweise, Merkmale für Proteininstabilitäten *in vivo* zu ermitteln, ist eine Analyse der Anzahl von Lysinresten. Lysinreste sind Stellen möglicher Ubiquitinierung und nachfolgender Degradation von Proteinen (zusammengefasst von Hochstrasser 1996). Es wurden zunächst die Lysinreste pro Protein betrachtet, sowohl als absolute Anzahl (Abb. 22A) als auch bezogen auf die Proteinelänge (Abb. 22B). Es konnte keine Korrelation zwischen der Anzahl der Lysinreste und der Expressionshöhe beobachtet werden.

Um die Zugänglichkeit der jeweiligen Reste für den Ubiquitinierungsapparat der Zelle zu berücksichtigen, wurde die Oberflächenwahrscheinlichkeit der Reste im Proteinmolekül berechnet und mit in Betracht gezogen. Abbildung 22C und D zeigen, dass es keine Korrelation zwischen Lysinresten mit hoher Oberflächenwahrscheinlichkeit (Emini-Index ≥ 1) und Expressionshöhe gibt, weder was die absolute Anzahl, noch was die Anzahl bezogen auf die Proteinelänge betrifft.

Die Anzahl der Lysinreste – weder absolut noch unter Berücksichtigung der jeweiligen Oberflächenwahrscheinlichkeit – zeigt eine Korrelation zur Expressionshöhe des Proteins.

4.3.3.4.3 Hydrophobe Bereiche

Stark hydrophobe Motive wurden ebenfalls als Teil von Signalen, die zur Ubiquitinierung und Degradation von Proteinen in *S. cerevisiae* führen können, beschrieben (Gilon et al. 2000). Darüber hinaus können diese Bereiche während der Proteinsynthese - bevor das synthetisierte Polypeptid lang genug ist, um sich zu falten und hydrophobe Bereiche im Molekülinneren „abzuschirmen“ – zur Aggregatbildung führen (Dobson und Karplus 1999). Proteinaggregate

werden in *S. cerevisiae* schneller degradiert als native Proteine (Saris et al. 1997).

Wie oben gezeigt (siehe 3.6.7), korreliert die Expressionshöhe nicht mit der generellen Hydrophobizität, diese ist jedoch ein Durchschnittswert über das gesamte Protein und spiegelt nicht lokale Häufungen hydrophober Reste wieder.

Aus diesen Gründen wurde das Vorhandensein hydrophober Bereiche in den Proteinen analysiert.

Wie in Abbildung 23 zu sehen ist, gibt es keinen signifikanten Zusammenhang zwischen hydrophoben Bereichen und der Expressionshöhe des jeweiligen Proteins.

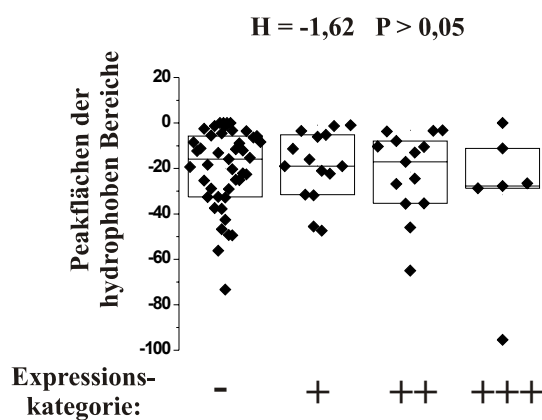


Abbildung 23: Verteilung der hydrophoben Bereiche in den Kategorien der Expressionshöhen

Ermittelt wurden die hydrophoben Bereiche. Aminosäuren mit einem Kyte-Doolittle Index von ≤ 1 wurden als Teil eines hydrophoben Bereiches angesehen.

4.3.3.5 Ähnlichkeiten der untersuchten Sequenzen zu annotierten *S. cerevisiae* Proteinen

Weiterhin stellte sich die Frage, ob Ähnlichkeiten der untersuchten humanen Proteine zu Hefeproteinen einen Einfluss auf die Expressionshöhe haben. Daher wurden zum einen Ähnlichkeiten der Proteinsequenzen zu *S. cerevisiae* Proteinen mittels BLAST quantifiziert, zum anderen wurde untersucht, wie oft Proteine derselben Strukturklasse wie die hier untersuchten in *S. cerevisiae* vorkommen.

4.3.3.5.1 Ähnlichkeiten zu Hefeproteinen aufgrund lokaler Sequenzalignments – Vergleich mittels BLAST

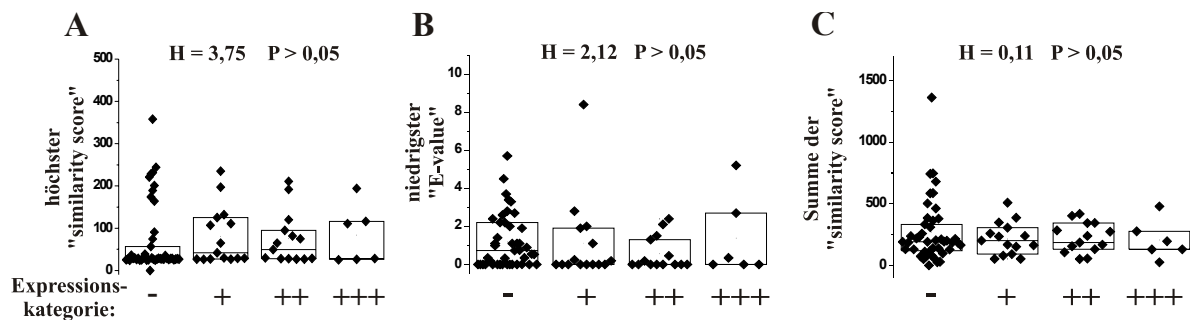


Abbildung 24: Quantifizierung der Ähnlichkeit der untersuchten Proteine zu *S. cerevisiae* Proteinen mittels BLAST

Für jede analysierte Proteinsequenz erzeugte BLAST mehrere lokale Alignments mit dem *S. cerevisiae* Genom. Zunächst wurden nur die besten Alignments jeder Sequenz ausgewertet. Aufgetragen wurde dazu der höchste „similarity score“ (A) sowie der niedrigste „E-value“ (B). Um zusätzlich die Anzahl der gefundenen Alignments zu berücksichtigen, wurde die Summe der „similarity scores“ pro Protein aufgetragen (C). Es ist keine Korrelation von einem dieser Parameter mit der Expressionshöhe zu sehen.

Bei Sequenzvergleichen mittels BLAST können sich für jede untersuchte Proteinsequenz mehrere lokale Alignments mit *S. cerevisiae* Proteinen ergeben. Jedem dieser Alignments ist ein „similarity score“ – die Summe der für die zugeordneten Reste vergebenen Ähnlichkeiten bezogen auf die Anzahl der zugeordneten Reste – und ein „Expect-value“ (E-value) – die Wahrscheinlichkeit, dass der erzielte „similarity score“ bei einer Suche gegen eine zufällige Datenbank von der benutzten Größe und Zusammensetzung erzielt wird – zugeordnet (Mount 2001). Aufgrund des niedrig gesetzten Schwellenwertes (E-value = 10) ergab die Analyse für alle Proteine bis auf Nr. 286 mindestens einen gefundenen ORF in *S. cerevisiae*. Zunächst wurde nur der jeweils beste Treffer berücksichtigt. In Abbildung 24A ist die Verteilung des höchsten gefundenen „similarity score“ der jeweiligen Proteinsequenz in den Kategorien der Expressionshöhe aufgetragen. In Abbildung 24B ist das zweite Kriterium für die Qualität eines lokalen Alignments, der „E-value“, ausgewertet. Der jeweils niedrigste „E-value“ pro Sequenz wurde aufgetragen.

Um zusätzlich zu ihrer Qualität die Anzahl der Treffer zu den *S. cerevisiae* ORFs mit einzubeziehen, wurden die gefundenen „similarity scores“ für jedes Protein addiert (Abb 24C).

Keiner der betrachteten Parameter zeigt eine Korrelation zu den Kategorien der Expressionshöhe.

4.3.3.5.2 Zuordnung zu strukturellen Protein-Superfamilien aus *S. cerevisiae*

Die Einordnung von unbekannten Proteinen mittels der SUPERFAMILY Datenbank basiert auf „hidden Markov models“ (HMMs). HMMs sind gemeinsame Sequenzprofile einer Gruppe von Sequenzen, die aufgrund von multiplen Alignments ermittelt wurden. Diese

Profile können genutzt werden, eine Datenbank oder einen Set von Proteinsequenzen auf Vertreter dieser Gruppe hin zu durchsuchen (Mount 2001). Im Unterschied zu anderen Proteindatenbanken, die auf HMMs basieren, z.B. Pfam (Bateman et al. 2000) oder SMART (Schultz et al. 2000), umfasst SUPERFAMILY nur Proteingruppen mit mindestens einem Vertreter bekannter Struktur (Gough und Chothia 2002). Die Zuordnung von Proteinen in SUPERFAMILY basiert gemeinsamen Strukturmerkmalen („structural classification of proteins“ (SCOP)). Dies führt zu einer Zuordnung der unbekannten Proteine zu Superfamilien aufgrund von Sequenzprofilen, die zu einer bestimmten Strukturgruppe gehören.

Es wurde sowohl die Anzahl der in *S. cerevisiae* gefundenen Proteine mit gemeinsamen SUPERFAMILY Motiven zu den analysierten Sequenzen – also die Anzahl der in *S. cerevisiae* vorkommenden Proteine aus der gleichen strukturellen Gruppe wie das analysierte Protein – (Abb. 25A), als auch die Anzahl der HMMs (die hier meistens Domänen repräsentieren) pro Protein, die in *S. cerevisiae* gefunden wurde (Abb. 25B), betrachtet. Diese beiden Werte sind unterschiedlich, da untersuchtes Protein und gefundenes *S. cerevisiae* Protein mehrere gemeinsame Domänen haben können.

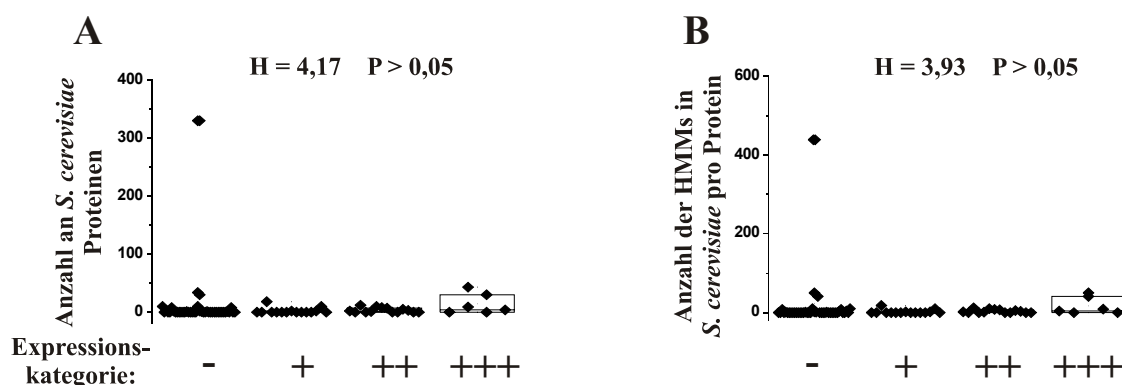


Abbildung 25: Zuordnung der Proteinsequenzen zu strukturellen Superfamilien

Die Proteinsequenzen wurden anhand der SUPERFAMILY Datenbank Protein-Superfamilien zugeordnet und es wurde bestimmt, wie oft diese Superfamilien im *S. cerevisiae* Proteom vorkommen. Teil A zeigt die Anzahl von Proteinen aus der entsprechenden Superfamilie in *S. cerevisiae* pro untersuchter humaner Sequenz. Da in den untersuchten Sequenzen in der Regel mehrere in der Datenbank vorhandene Sequenzmotive – hier: HMMs – vorhanden sind, und diese auch mehrfach in einem *S. cerevisiae* Protein vorkommen können, wurde weiterhin die Anzahl der HMMs pro Protein, die im *S. cerevisiae* Proteom vorkommen, aufgetragen (B).

Ein Zusammenhang zwischen der Ähnlichkeit der Sequenzen zu putativen *S. cerevisiae* Proteinen und der Expressionshöhe in *P. pastoris* ist nicht feststellbar.

4.3.3.6 Sekundärstrukturmerkmale nach PSIPRED und die Expressionshöhe

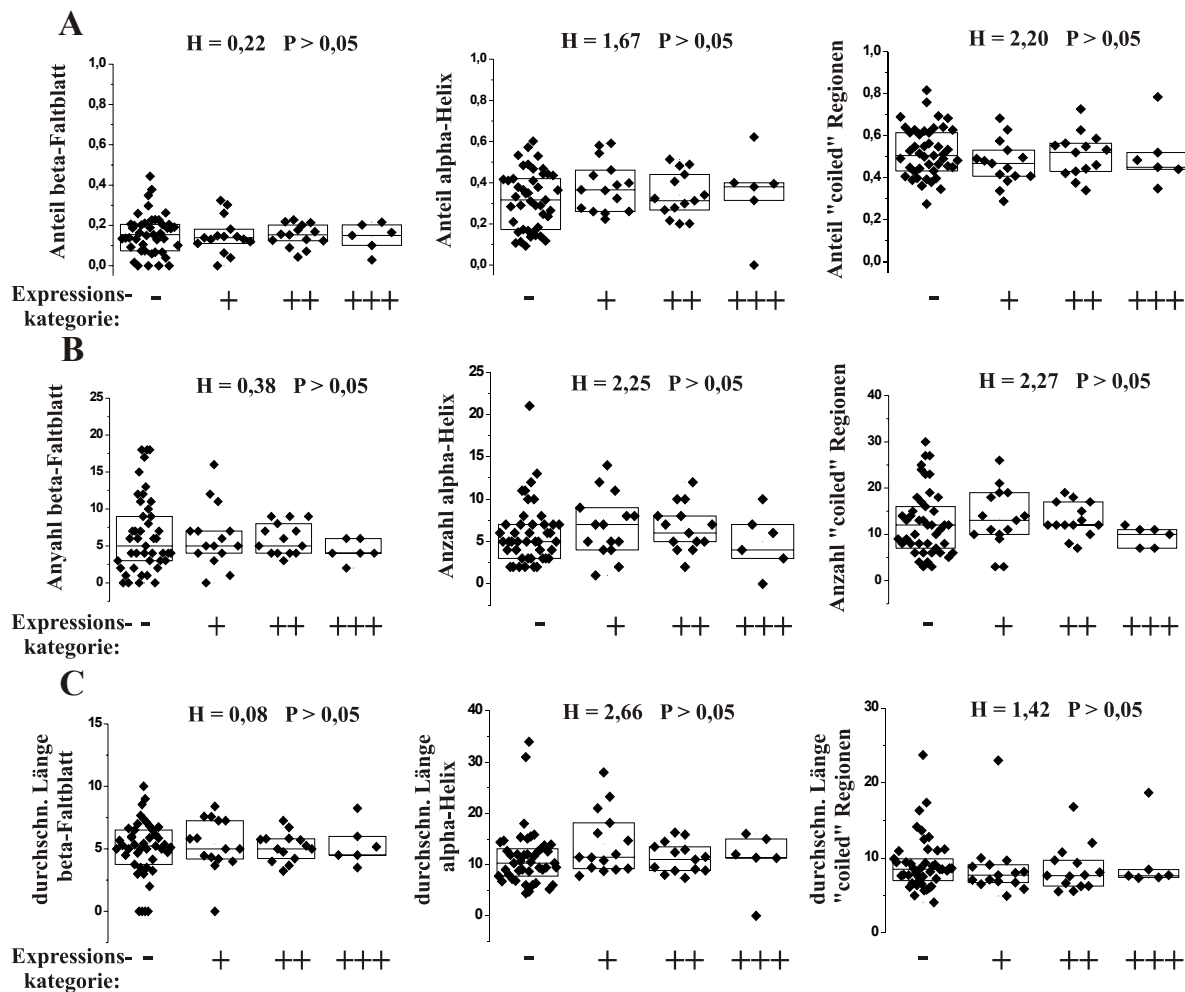


Abbildung 26: Verteilung von Sekundärstrukturmerkmalen

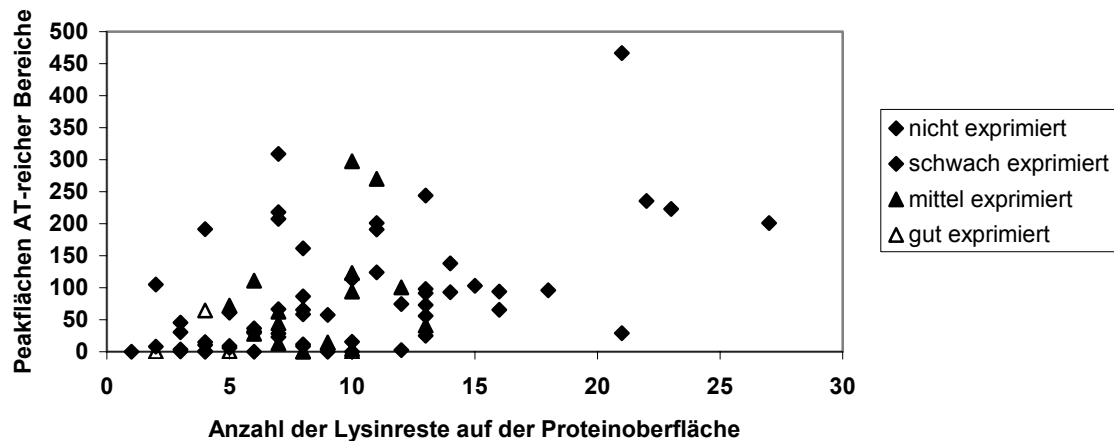
Analysiert wurde der Anteil der Sekundärstrukturmerkmale Alpha-Helix, Beta-Faltblatt und „coiled“-Regionen, wobei als „coiled“-Region bezeichnet wird, was nicht in die ersten beiden Kategorien fällt. Aufgetragen wurden die Anteile der Merkmale an der Proteinsequenz (A), die Anzahl der Merkmale in der Sequenz (B) und die durchschnittliche Länge der Merkmale in dem jeweiligen Protein (D). Keines dieser Merkmale korreliert mit der Expressionshöhe.

Analysiert wurde, ob Sekundärstrukturmerkmale mit der Expressionshöhe korrelieren. Das Programm PSIPRED wurde zur Vorhersage der Sekundärstrukturmerkmale gewählt, weil es auf einer Methode beruht, die in einem Vergleich verschiedener Vorhersagen als eine der beiden genauesten beschrieben worden ist (McGuffin und Jones 2003). Das Programm ordnet jedem Aminosäurerest der Sequenz ein Sekundärstrukturmerkmal zu. Differenziert wird dabei zwischen α -Helices, β -Faltblätter sowie Regionen, die nicht zu diesen Gruppen gehören („coils“). Betrachtet wurde der jeweilige Anteil der Sekundärstrukturmerkmale an der Gesamtsequenz des Proteins (Abb. 26A), die Anzahl der zusammenhängenden Sequenzen gleicher Zuordnung, (Abb. 26B) und die durchschnittliche Länge der Merkmale im Protein (Abb. 26C).

Keiner dieser Parameter korreliert mit einer der Kategorien der Expressionshöhe. Bestimmte Sekundärstrukturmerkmale bzw. deren Anteil im Protein scheinen keinen Einfluss auf die Expressionshöhe zu haben.

4.3.4 Zweidimensionale Auswertungen – Kombination verschiedener Parameter

A



B

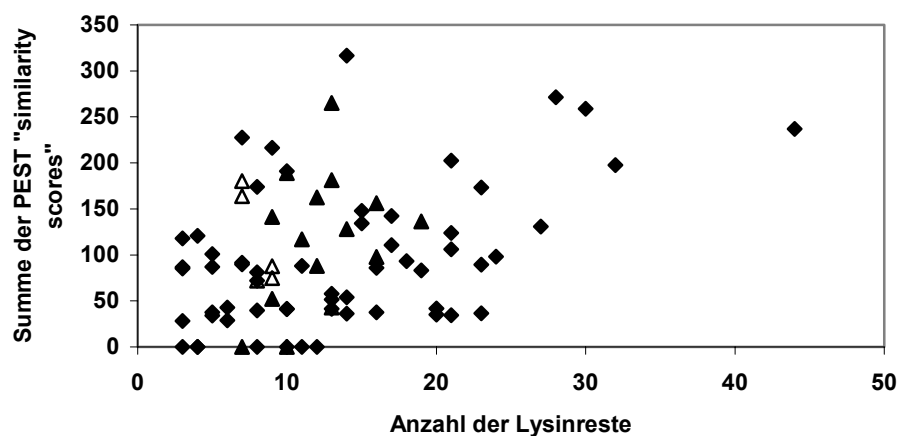


Abbildung 27: Die zweidimensionale Analyse ausgewählter Merkmalskombinationen

Aufgetragen ist jeweils ein Parameter pro Achse. Bei additiven Effekten der aufgetragenen Merkmale müssten in bestimmten Bereichen des Plots bevorzugt Klone einer Kategorie der Expressionshöhe auftreten.

Bei heterologer Expression in *P. pastoris* kann es zu additiven Effekten beim Auftreten mehrerer nachteiliger Parameter kommen (Sinclair und Choy 2002). Aus diesem Grund wurde die Kombination von Parametern in den jeweiligen Sequenzen analysiert. Bei den hierzu erstellten zweidimensionalen Plots ist jeweils ein Parameter auf jeder Achse aufgetragen. Bei Merkmalen, die alleine eine Expression z.B. nicht ausschließen, in

Kombination mit einem anderen Parameter sich jedoch deutlich nachteilig auswirken, müsste in einer solchen Darstellung eine Anhäufung negativer Klone in einem bestimmten Bereich des Diagrammes auftreten. Es wurden bevorzugt Parameter in diese Form der Auftragung einbezogen, bei denen deutliche Ausreißer in einer Kategorie auftraten oder deren Verteilung innerhalb einer Kategorie ungleichmäßig war.

Dies waren die Anzahl der effektiven Codone, die Anzahl seltener Codone, die Proteinelänge, die Summe der PEST „similarity scores“, der höchste PEST „similarity score“ bezogen auf die Proteinelänge, die Anzahl der Lysinreste bezogen auf die Proteinelänge, die Anzahl der Lysinreste mit einer berechneten Lokalisation an der Proteinoberfläche und die Summe der BLAST „similarity scores“. Diese Merkmale wurden jeweils mit den beiden Parametern kombiniert, für die eine Korrelation mit der Expressionshöhe gefunden wurde (Peakfläche der AT-reichen Bereiche und isoelektrischer Punkt der Proteine).

Außerdem wurden die Merkmale, die im Zusammenhang mit einer Proteindegradation betrachtet wurden, kombiniert.

Für die Interpretation der Diagramme ist wichtig, dass die Häufungen von Klonen bestimmter Kategorien sich durch Diagonale von den übrigen Punkten abtrennen lassen müssen, verlaufen die Grenzen zwischen den Kategorien senkrecht oder vertikal sind die in der Boxplotdarstellung bereits erkennbaren Ausreißer die Ursache und nicht ein additiver Effekt der Parameter. Ein Beispiel für eine solche Diagonale zeigt Abbildung 27A. Im rechten oberen Bereich sieht man ausschließlich nicht exprimierte Sequenzen. Das könnte ein Hinweis auf einen additiven Effekt sein, allerdings ist die Zahl der Sequenzen im fraglichen Bereich klein.

Teil B zeigt die Kombination aus der Summe der PEST „similarity scores“ und der Anzahl der Lysinreste. Hier ist es schwierig zu beurteilen, ob im rechten oberen Bereich eine Häufung von nicht exprimierten Sequenzmerkmalen auftritt oder ob die Lage der Punkte daraus resultiert, dass nur in der Kategorie der nicht exprimierten Proteine Sequenzen mit hohem Lysingehalt auftreten.

Diese Auswertungen können nicht auf statistische Signifikanz getestet werden und müssen entsprechend vorsichtig beurteilt werden. Deshalb und aufgrund der Anzahl der Darstellungen werden nur ausgewählte Beispiele abgebildet. In den nicht abgebildeten Plots sind keine Häufungen von einzelnen Kategorien der Expressionshöhe zu sehen.

5 Diskussion

Die vorliegende Arbeit ist in zwei größere Abschnitte geteilt. Im ersten wurden Methoden zur heterologen Expression von humanen Proteinen in der Hefe *P. pastoris* im Rahmen von Hochdurchsatzverfahren entwickelt. Hierzu wurde ein parallelisiertes „screening“-System in kleinem Maßstab und soweit als möglich automatisierbar zur Erstellung und Charakterisierung von Expressionsklonen unter standardisierten Bedingungen aufgestellt und angewandt. In dem zweiten Teil wurde mit der damit erstellten Sammlung von Expressionsklonen gearbeitet, um Erkenntnisse über Zusammenhänge zwischen Sequenz der cDNA und Expressionshöhe zu gewinnen.

5.1 Entwicklung eines parallelisierten Systems zum „screening“ von Expressionsklonen und Erstellung einer Sammlung von Expressionsklonen

Eine wichtige Limitierung von Hochdurchsatzprojekten zur Bestimmung von Proteinstrukturen besteht in der heterologen Expression der Proteine (Edwards et al. 2000; Vincentelli et al. 2003; Yokoyama 2003). Die Entwicklung von Hochdurchsatzverfahren zur Erstellung von Expressionsklonen für „structural genomics“ umfasst die Klonierung der cDNAs in Expressionsvektoren sowie die Transformation der resultierenden Konstrukte und die folgende Charakterisierung der Expressionsklone für Hunderte von Zielproteinen (Yokoyama 2003). Um dieses zu bewerkstelligen, müssen die notwendigen Teilschritte weitestgehend parallelisiert und miniaturisiert werden. Diese Verfahren sind etabliert für die Expressionswirte *E. coli* (z.B. Christendat et al. 2000; Gilbert und Albala 2002; Doyle et al. 2003; Scheich et al. 2003; Vincentelli et al. 2003) und *S. cerevisiae* (Holz et al. 2003) sowie für Zell-freie Proteinexpression (z.B. Sawasaki et al. 2002; Yokoyama 2003). Im Rahmen der vorliegenden Arbeit wurde ein entsprechendes System für die Hefe *P. pastoris* entwickelt (Boettner et al. 2002).

5.1.1 Variation von Zufütterung und Medium

Um mit *P. pastoris* in dem zur Parallelisierung notwendigen kleinen Kultivierungsmaßstab ein möglichst niedriges Detektionslimit für die Proteinexpression zu erzielen, musste zunächst das Medium und die Induktion des Promotors im Schüttelkolben in einem 50 mL Maßstab optimiert werden.

Bei der Nutzung des *AOX1*-Promotors für die Expression in *P. pastoris* ist Methanol sowohl Induktor als auch C-Quelle (Ellis et al. 1985). Dies ist bedingt durch die Repression des Promotors durch andere C-Quellen wie Glycerin oder Glukose (Tschopp et al. 1987). Um toxische Effekte durch zu hohe Methanolkonzentrationen und folgende Akkumulation der Metaboliten Formaldehyd und Wasserstoffperoxid (Couderc und Baratti 1980) zu vermeiden und gleichzeitig eine gute Induktion des Promotors zu erzielen, ist es wichtig, die Methanolkonzentration innerhalb eines engen Rahmens zu halten (Guarna et al. 1997). So wurde für eine Methanol-limitierte Fermentation eine maximale Produktausbeute bei intrazellulärer Expression bei einem Drittel der maximalen Wachstumsrate beobachtet (Zhang et al. 2000). Für optimale Ausbeuten mit geregelter Methanolkonzentration existiert für Bioreaktoren eine entsprechende Regelungstechnik bzw. Protokolle (Hellwig et al. 2001; Prinz et al. eingereicht). Ein Ziel dieser Arbeit war es, für den Schüttelkolbenmaßstab empirisch eine Zufütterung von Methanol für die hier benutzten Expressionsstämme zu entwickeln, die ohne Mess- und Regeltechnik auskommt und parallelisierbar ist. Die meisten in der Literatur beschriebenen Induktionsprotokolle (z.B. Berrin et al. 2000; Moreno et al. 2000; Newton-Vinson et al. 2000; Zhou et al. 2000) basieren auf den Empfehlungen von Invitrogen™, CA (Invitrogen 1997). Induziert wird dabei durch Zugabe von Methanol auf eine Endkonzentration von 0,5 % (v/v). Diese Zugabe wird nach 24 h wiederholt.

Bei den hier durchgeführten Versuchen bestätigte sich die Abhängigkeit der Expressionshöhe von der optimalen Konzentration des Induktors Methanol. Bei Variation der zugeführten Methanolmenge ändert sich die Proteinexpression stärker als das Wachstum. Eine Steigerung der Methanolmenge um den Faktor drei (von 1,0 % (v/v) Endkonzentration zweimal täglich auf 3,0 %) senkt die Proteinexpression um den Faktor 2,4, hat jedoch keinen Einfluss auf das Zellwachstum. Die vergleichbaren Wachstumsraten zeigen, dass die Abnahme der Expression nicht durch Änderungen der Energieversorgung der Zellen oder toxische Effekte bedingt durch die höhere Methanolkonzentration (Couderc und Baratti 1980) verursacht ist. Hier spielt vielmehr die optimale Konzentration des Induktors eine Rolle, z.B. aufgrund einer „feedback“-Hemmung durch Intermediate des Methanol Stoffwechsels. Vergleichbares wurde bereits für Fermentationen beobachtet (Zhang et al. 2000).

Die geringere Expressionshöhe bei 0,5 % (v/v) Methanol spiegelt sich in der Wachstumsrate wieder, hier liegt ein Mangel an Methanol vor. Das verringerte Wachstum bei hohen Methanolkonzentrationen deutet auf toxische Effekte hin. Bei einer Zufütterung zweimal täglich im Schüttelkolben ist 1,0 % (v/v) Methanol also die optimale Endkonzentration. Hier

konnte eine Steigerung der Expressionshöhe um den Faktor 2,4 gegenüber dem üblichen Protokoll erreicht werden.

Da der optimale Bereich der Methanolkonzentration relativ eng ist, führt eine diskontinuierliche Fütterung zu einer suboptimalen Methanolkonzentration über weite Teile der Expressionsphase. Das führt zu verminderten Proteinausbeuten (Guarna et al. 1997). Im Hinblick auf einen vertretbaren manuellen Aufwand können diese Verluste jedoch in Kauf genommen werden, zumal gezeigt wurde, dass das erzielte Detektionslimit für die im Rahmen der PSF interessante Expressionshöhe ausreichend ist.

Hefen säuern das Medium während des Wachstums stark an, als Folge fällt der intrazelluläre pH-Wert ebenfalls ab (Slavik und Kotyk 1984). Da humane Proteine zur Strukturaufklärung exprimiert werden sollten, war ein neutraler pH im Cytosol wichtig, um eine native Konformation der Zielproteine nicht zu stören. Aus diesem Grunde wurde nach dem Vergleich verschiedener Medien trotz geringfügig höherer GFP-Expression in WM8 das gepufferte WM9-Medium genutzt.

Gemischte Zufütterungen während der Induktionsphase wurden in Fermentationen von Mut⁺-Stämmen wiederholt erfolgreich angewandt unter Zusatz von Glycerin als zusätzlicher C-Quelle (McGrew et al. 1997; Hellwig et al. 2001; Zhang et al. 2003). McGrew und Mitarbeiter (1997) erreichten die höchsten Proteinerträge in einer Fermentation bei einem Glycerin/Methanol Verhältnis von 1:1. Die Zufütterungsraten während der Induktionsphase wurden durch den gelösten Sauerstoff gesteuert, so dass die Rate für Glycerin aus den publizierten Daten nicht ermittelbar ist. Hellwig und Mitarbeiter (2001) fanden, ebenfalls während einer Fermentation, bei einer konstanten Glycerinzufütterung von $1,23 \text{ g} \cdot \text{l}^{-1} \cdot \text{h}^{-1}$ bei konstanten 0,5 % Methanol im Medium einen reprimierenden Effekt auf die Proteinproduktion. Die in der vorliegenden Arbeit als optimal ermittelte Zufütterung von 0,1 % Glukose zweimal täglich entspricht einer Zufütterungsrate von $0,083 \text{ g} \cdot \text{l}^{-1} \cdot \text{h}^{-1}$. Selbst unter Berücksichtigung der Tatsache, dass Glukose ein C₆-Körper ist und möglicherweise anders verwertet wird als Glycerin sowie im Schüttelkolben geringere Zelldichten erreicht werden als im Bioreaktor, liegt hier wahrscheinlich eine deutlich geringere spezifische Konzentration an reprimierender C-Quelle vor als in den genannten Arbeiten. Diese Ergebnisse zeigen, dass auch im Schüttelkolben eine verbesserte Produktivität durch eine gemischte Zufütterung zu erreichen ist. Das Endniveau der GFP-spezifischen Fluoreszenz bei optimaler gemischter Zufütterung entspricht dem von ausschließlicher Methanol-Induktion. Dieses Endniveau wird bei gemischter Zufütterung allerdings schneller erreicht. Die schnellere Zunahme der Proteinmenge spiegelt sich in einem schnelleren Wachstum der

Zellen. Das deutet darauf hin, dass die Zunahme der Proteinmenge durch die Zunahme der Zellzahl bedingt ist.

5.1.2 Zusammenhang zwischen Integration der Expressionskassette und Proteinexpression

In der Literatur ist auf die Notwendigkeit, Transformanten auf DNA-Ebene zu überprüfen, hingewiesen (Invitrogen 1997; Andrin et al. 2000; Bisht et al. 2001). Als Vorarbeit zur Entwicklung des „screening“-Systems musste geklärt werden, wie viele Histidin-prototrophe Kolonien, die mit einem Vektor mit exprimierbarer cDNA transformiert wurden, falsch positiv sind, also kein Protein exprimieren. Für die Entwicklung eines Hochdurchsatzverfahrens zur Erstellung und Charakterisierung von Expressionsklonen ist es essentiell zu wissen, wie viele individuelle Transformanten pro klonierter cDNA überprüft werden müssen. Von den per Kolonie-PCR überprüften His⁺-Klonen waren 69 % positiv für die Expressionskassette, wovon lediglich ein Klon das Reporterprotein GFP nicht exprimierte. Eine Überprüfung der Transformanten auf DNA-Ebene ist somit notwendig, da bei 31 % der His⁺-Kolonien weder die Expressionskassette noch Protein detektiert werden konnte. Die gute Korrelation von PCR-Ergebnis und Proteinexpression zeigt, dass die Überprüfung von zwei PCR-positiven Klonen auf Proteinexpression ausreicht, um mit sehr hoher Wahrscheinlichkeit falsch negative Ergebnisse in Bezug auf die Exprimierbarkeit einer cDNA zu vermeiden.

Ein negatives PCR-Ergebnis ist wahrscheinlich begründet in einem Rekombinationsereignis während der Transformation, das einen funktionalen *HIS4*-Lokus erzeugt, ohne Integration der Expressionskassette. Eine Reversion des auxotrophen Phänotyps in Höhe der beobachteten Rate ist unwahrscheinlich, da Kontroll-Transformationen ohne eingesetztem Plasmid unter Standard Bedingungen in keinem Fall His⁺-Kolonien erzeugten.

5.1.3 Entwicklung der Expressionskontrolle im 2 mL Maßstab

Die Entwicklung und Optimierung von Hochdurchsatzprojekten erfordert die Identifizierung der geschwindigkeitsbestimmenden Schritte im Arbeitsablauf, um diese gezielt zu beschleunigen (Christendat et al. 2000). Für Proteinexpressionsprojekte wurden für die Expression in *E. coli* die Klonierungen sowie die Vorbereitung und Durchführung der Protein-Gelelektrophoresen als limitierend erkannt (Hammarstroem et al. 2002). Im vorliegenden Falle stellte sich die Frage, wie weit die Expression in *P. pastoris* miniaturisiert werden muss, um nicht der limitierende Schritt zu sein. Aufgrund des hohen Sauerstoffbedarfs bei Verstoffwechselung von Methanol im Vergleich zu anderen C-Quellen

(Cereghino und Cregg 2000) ist ein hohes Oberflächen zu Volumen Verhältnis der Kultur wichtig, um den notwendigen Sauerstoffeintrag zu gewährleisten. Aus diesen Gründen wurde eine Reduzierung der Kultivierung auf den 2 mL Maßstab in 24-„well“ Platten entwickelt. Es zeigte sich, dass nach der Verkleinerung des Expressionsmaßstabes die limitierenden Schritte des Prozesses an anderer Stelle lagen. Zusätzlich zu den oben genannten Punkten ist in diesem Zusammenhang noch die Transformation der Hefen mittels Elektroporation zu nennen. Die am häufigsten genutzten Transformationsmethoden für *P. pastoris* sind Elektroporation (z.B. Maeda et al. 2000; Moreno et al. 2000; Bisht et al. 2001; Li et al. 2001) sowie Spheroplastentransformation (z.B. Fahnestock und Bedzyk 1997; Andrin et al. 2000; Berrin et al. 2000; Newton-Vinson et al. 2000). Eine weniger genutzte Methode ist die Transformation chemisch kompetenter Zellen (z.B. Mwangi et al. 2000; Mochizuki et al. 2001). Die in der vorliegenden Arbeit erreichten Transformationsraten mit chemisch kompetenten Zellen lagen zu niedrig, um mit der Ausbeute einer „Mini-präp“ Plasmidisolation zu transformieren. Da elektrokompetente Zellen bei -70°C lagerbar sind (Invitrogen 1997), wurde hier mittels Elektroporation transformiert.

5.2 Ursachen für die beobachteten Unterschiede in der Expressionshöhe

Nach der erfolgten Expressionsanalyse der cDNAs in *P. pastoris*, wurden die Klone auf Eigenschaften untersucht, die Ursache für die beobachteten unterschiedlichen Expressionshöhen sein könnten. Hierzu wurden zunächst die Mengen an entsprechender mRNA sowie deren Stabilitäten in verschiedenen Expressionsklonen verglichen.

Weiterhin wurden mittels bioinformatischer Methoden Parameter identifiziert, die mit den beobachteten Expressionshöhen korrelieren.

5.2.1 Menge und Stabilität der Transkripte

Gygi und Mitarbeiter (1999) stellten bei der Analyse von 150 *S. cerevisiae* Genen fest, dass das „steady state“ Niveau von Transkripten nicht mit der Expressionshöhe der entsprechenden Proteine korreliert. Die Analyse der Transkriptmengen ausgewählter *P. pastoris* Expressionsklone nach 1,5 Stunden zeigte ebenfalls keinen Zusammenhang zwischen der Menge an Transkript und Höhe der Proteinexpression. Das „steady state“ Niveau von mRNAs scheint also kein gutes Kriterium für die Expressionshöhe der entsprechenden Proteine zu sein.

Hingegen wurde die Stabilität von Transkripten als ein wichtiges Element der Genregulation beschrieben (zusammengefasst von Mitchell und Tollervey 2000). In *S. cerevisiae* variiert die

Halbwertszeit von endogenen Transkripten stark zwischen $t_{1/2} > 25$ min und $t_{1/2} < 7$ min, wobei Transkriptstabilitäten in einem Bereich von $t_{1/2} < 7$ min als gering eingestuft werden (Herrick et al. 1990). Daten zur Stabilität von Transkripten von einer *AOX1*-Expressionskassette in *P. pastoris* sind in der Literatur nicht vorhanden.

Mit der hier durchgeführten Methode der Promotorrepression durch Glukose und Messung der Transkriptabnahme konnten keine genauen Halbwertszeitmessungen durchgeführt werden. Dies lag an der geringen Stabilität der Transkripte in Verbindung mit der Notwendigkeit vor der Analyse einen Mediumwechsel vornehmen zu müssen. Aus Zeitgründen konnten in dieser Arbeit keine weiteren Versuche zur Hemmung der Transkription durch z.B. Thiolutin (Das et al. 2000) durchgeführt werden.

Alle hier überprüften Transkripte haben unabhängig von der Expressionshöhe der kodierten Proteine eine Halbwertszeit von weniger als fünf Minuten (043 und 048) bzw. zehn Minuten (GFP). In Eukaryonten korrelieren kurze Halbwertszeiten mit dem Vorkommen AU-reicher Bereiche (AREs) in 3'-UTRs (Chen und Shyu 1995). Solche Bereiche sind in der 3'-UTR, die durch den im Expressionsvektor pPICH5 vorhandenen AOX1-Terminator zustande kommen, ebenfalls vorhanden (siehe Abbildung 28 im Anhang).

Interessant ist die Abnahme der Transkriptmenge nach Umsetzen der Zellen auf frisches WM9 Medium mit Glutamat als N-Quelle. Als reprimierend für den *AOX1*-Promotor sind Glukose und Glycerin beschrieben (Tschopp et al. 1987). Das Kohlenstoffgerüst von Glutamat kann nach Desaminierung über α -Ketoglutarat in den Citratzyklus eingehen (Stryer 1991) und somit zur Bildung kataboler Metaboliten beitragen, durch die eine Repression C-Quellen abhängiger Promotoren denkbar wäre.

Aus der Inhibierung des Promotors durch Glutamat oder dessen Stoffwechselprodukten ergibt sich eine mögliche Verbesserung des Expressionsmediums. Ein Austausch der N-Quelle zu NH_4Cl resultiert in einer stabilen Transkriptmenge nach Wechsel auf frisches Medium. Da eine Voraussetzung für eine vergleichende Analyse der Sequenzen auf Determinanten für den Expressionserfolg eine standardisierte Charakterisierung der Klone ist, wurde diese Erkenntnis innerhalb dieser Arbeit nicht umgesetzt.

Einen ersten Hinweis auf Unterschiede in den Stabilitäten der verschiedenen Transkripte könnten die verschiedenen Abnahmeraten der mRNAs von Klon 048 (gute Expression) und Klon 043 (keine Expression) unter Induktionsbedingungen in frischem glutamathaltigem Medium geben. Hier nimmt die mRNA von Klon 048 deutlich langsamer ab als die des Klons 043. Aufgrund der induzierenden Bedingungen in diesem Experiment ist jedoch keine

Halbwertszeitbestimmung möglich, man sieht ein Gleichgewicht zwischen Synthese- und Abbauraten. Es ist nicht auszuschließen, dass die Syntheseraten unterschiedlich sind.

5.2.2 Die Qualität der Sequenzinformation

Bei jeder sequenzbasierten Analyse muss nach der Qualität der zugrundeliegenden Sequenzinformation gefragt werden. Die im Rahmen der PSF verwandten cDNA-Ausgangsklone wurden mittels Ansequenzierung des 5'-Endes der klonierten Sequenz identifiziert. Dabei wurden nicht alle klonierten cDNAs durchsequenziert, so dass keine vollständigen bestätigten Sequenzinformationen vorliegen. Die erste PCR, die innerhalb der PSF von diesen Klonen durchgeführt wurde um die cDNAs mit den Restriktionschnittstellen zur Klonierung zu fusionieren, erfolgte mit genspezifischen Primern. Die Tatsache, dass mit diesen Primern ein PCR-Produkt entstand, zeigt dass auch das 3'-Ende der cDNAs den Erwartungen entsprach. Die Größe der Amplifikate zeigt – innerhalb der mit Agarosegelen zu erzielenden Auflösung –, dass keine alternativen Spleißprodukte kloniert wurden.

Punktmutationen dürften im Verlauf der Erzeugung der cDNA und der Umklonierungen in die verschiedenen Expressionsvektoren vor allem durch die Nutzung von reverser Transkriptasen zur Generierung der cDNAs innerhalb des I.M.A.G.E. Konsortiums eingeführt worden sein, da diese Enzyme über keine korrigierenden Exonukleaseaktivitäten verfügen (Roberts et al. 1988). Die Fehlerrate der verschiedenen reversen Transkriptasen beträgt zwischen 1:4000 und 1:30000 (Preston et al. 1988; Roberts et al. 1988; Roberts et al. 1989; Manns et al. 1991). Sequenzanalysen wie z.B. Alignments liefern noch sinnvolle Ergebnisse bei Fehlern in den Sequenzdaten von 5 % Basensubstitutionen oder 1 % Leserasterverschiebungen (States und Botstein 1991). Diese zulässigen Fehler liegen deutlich über den zu erwartenden Abweichungen von den theoretischen Sequenzen.

Als zusätzlicher Hinweis auf die Qualität der cDNAs ist die Exprimierbarkeit in den anderen Wirtsorganismen der PSF zu berücksichtigen. Der Anteil an exprimierten Proteinen in *S. cerevisiae* und *E. coli* lag jeweils bei ca. zwei Dritteln. Die exprimierbaren Proteine in diesen Organismen sind nicht immer identisch. (Dr. C. Holz, Dr. K. Büsow und C. Scheich, persönliche Mitteilung; siehe auch die öffentliche Statusseite der PSF unter www.proteinstrukturfabrik.de). Weiterhin wurde in dieser Arbeit kein rekombinantes Produkt beobachtet, dass sich nur über einen der beiden Tags nachweisen ließ, dessen Translation also frühzeitig abgebrochen wurde. Das zeigt, dass der größte Teil der klonierten Sequenzen kein Stopcodon innerhalb der Sequenz enthält.

Zusammengenommen sprechen diese Punkte für eine gute Qualität der Sequenzinformationen. Einzelne Punktmutationen oder Abweichungen zur Datenbanksequenz können nicht ausgeschlossen werden, in Anbetracht der Zahl der analysierten Sequenzen und der Art der analysierten Merkmale dürften diese jedoch keine deutlichen Auswirkungen auf das Ergebnis zeigen.

5.2.3 Ungerichtete Suche nach Sequenzmerkmalen, die mit der Expressionshöhe zusammenhängen

Um unbekannte sequenzbasierte Parameter zu identifizieren, werden ungerichtete Ansätze verfolgt. Hierbei werden Gemeinsamkeiten innerhalb eines ausgewählten Satzes von Sequenzen ohne eine *a priori* Annahme bezüglich des gesuchten Merkmals gesucht. So wurde bereits 1976 das Transkriptionsterminationsmotiv AAUAAA als eine Gemeinsamkeit von 3'-UTRs höherer Eukaryonten beschrieben (Proudfoot und Brownlee 1976). Für verschiedene Spezies wurden durch Vergleiche nicht-codierender DNA-Bereiche konservierte Motive gefunden, die Transkriptionsfaktoren binden (Hughes et al. 2000; Levy et al. 2001; Kellis et al. 2003). Ohne Nutzung vorhandener Kenntnisse über transkriptionell regulierende Bereiche konnten in *S. cerevisiae* Regulons anhand von Gemeinsamkeiten in den Promotorbereichen der entsprechenden Gene identifiziert werden (Tavazoie et al. 1999). Der Vergleich der genomischen Sequenz von *S. cerevisiae* mit der von zwei verwandten *Saccharomyces* Arten führte zusätzlich zur Beschreibung neuer putativer regulatorischer Sequenzen auch zu einer verbesserten Annotation von codierenden Sequenzen (Kellis et al. 2003). Auf Proteinebene konnte durch einen Vergleich von Proteinen aus thermophilen Mikroorganismen mit Orthologen aus mesophilen Mikroorganismen Mutationen identifiziert werden, die zu einer Thermostabilität von Proteinen führen (La et al. 2003). In den genannten Arbeiten wurden sowohl Sequenzalignments (Kellis et al. 2003; La et al. 2003) als auch Algorithmen zur Motivsuche (Tavazoie et al. 1999; Hughes et al. 2000; Levy et al. 2001) genutzt.

Beide Ansätze wurden in der vorliegenden Arbeit verfolgt. Zunächst wurde mittels globaler Alignments untersucht, ob die Expressionshöhe mit bestimmten Sequenzmerkmalen zusammenhängt. Hierzu wurden phylogenetische Bäume der Sequenzen erstellt. Sollten globale Gemeinsamkeiten in den Sequenzen entscheidend für die Expression sein, würden die entsprechenden Sequenzen sich anhand der Expressionshöhe im Baum anordnen. Entscheidend ist hier, ob die benutzten Algorithmen zur Detektion solcher Gemeinsamkeiten bzw. Unterschiede in der Lage sind. Grundsätzlich sind phylogenetische Analysen dazu

angelegt, den Verwandtschaftsgrad homologer Sequenzen zu bestimmen. Die der Konstruktion des Baumes zugrundeliegenden multiplen Alignments sind fehleranfällig bei geringen Sequenzähnlichkeiten (Mount 2001). Dies ist insbesondere dann der Fall, wenn die Sequenzen lange Bereiche enthalten, die nicht vergleichbar sind (Higgins et al. 1996). Im Unterschied zu den oben genannten Arbeiten von Kellis und Mitarbeitern (2003) und La und Mitarbeitern (2003) sind die hier analysierten Sequenzen nicht homolog und enthalten aufgrund ihrer großen Unterschiede große, nicht vergleichbare Bereiche. Trotz dieser prinzipiellen Schwierigkeit wurde untersucht, ob sich globale Gemeinsamkeiten den Expressionshöhen zuordnen lassen.

Gewählt wurde der Algorithmus ClustalW, da dieser – im Gegensatz zu vielen anderen – erstens relativ unempfindlich ist gegen ungleichmäßig verteilte Variationen innerhalb des Datensatzes (einige Sequenzen sind unterschiedlicher als andere) und zweitens in der Lage ist, eine Zahl von Sequenzen zu verarbeiten, die deutlich über zehn liegt (Higgins et al. 1996; Mount 2001). Die Autoren empfehlen aufgrund des großen möglichen Parameterraumes von ClustalW, das Programm eher als Mittel zur Untersuchung denn als exakte Analyse zu betrachten, also die Parameter empirisch zu variieren und die Ergebnisse anschließend auf Plausibilität zu überprüfen (Higgins et al. 1996). Entsprechend wurde hier verfahren. Es wurden jedoch unter keiner der gewählten Bedingungen die Sequenzen analog ihrer Expressionshöhe angeordnet. Dies kann verschiedene Ursachen haben. Die hauptsächlichen Limitationen des Ansatzes sind die oben beschriebenen: die geringe Ähnlichkeit der Sequenzen macht sinnvolle Alignments schwierig. Weiterhin ist die Funktionsweise von ClustalW zu nennen: als progressiver Alignment-Algorithmus geht dieser von einem Initialalignment der ähnlichsten Sequenzen im Datensatz aus und fügt die übrigen entsprechend ihrer Ähnlichkeit zum Ausgangsalignment hinzu (Higgins und Sharp 1988). Ähnliche Bereiche werden also als solche bevorzugt erkannt, wenn sie in den ähnlichsten Sequenzen des Datensatzes vorhanden sind bzw. Ähnlichkeiten zwischen Sequenzen, die nicht im Ausgangsalignment vorhanden sind, werden nicht erkannt. In diesem Datensatz ist es nicht notwendigerweise der Fall, dass die gesuchten gemeinsamen Merkmale in den Ausgangsalignmenten bereits als solche erkannt werden.

Darüber hinaus ist es möglich, dass die Expressionshöhe nicht durch globale Sequenzähnlichkeiten bestimmt wird.

Der zweite, häufig genutzte Ansatz neben der Suche nach globalen Gemeinsamkeiten ist die Analyse lokal begrenzter Sequenzmerkmale – sprich: Motive. Das Programm MEME (Bailey 1994) findet gemeinsame Motive in einem Datensatz, auch wenn diese nicht in allen

Sequenzen vorkommen. Dabei werden keine Sequenzen übergewichtet (Grundy et al. 1997), wie dies z.B. bei ClustalW durch das progressive Vorgehen geschieht. Die Autoren des Programms zeigten, dass MEME gemeinsame Motive findet, wenn diese nur in 20 % der Sequenzen des Datensatzes vorhanden sind (Bailey 1994). In der vorliegenden Arbeit erwies sich das Programm als deutlich empfindlicher. Von den gefundenen Motiven wurden nur die elf näher analysiert, die lediglich in einer Kategorie der Expressionshöhe vorkamen. Wie die SwissProt Annotation der Proteine zeigte, charakterisierten diese Motive bestimmte biologisch relevante Klassen wie z.B. Transmembranproteine mit vier Transmembranhelices oder Adenylyl-Zyklasten. Diese biologische Relevanz ist ein deutliches Indiz dafür, dass diese Motive signifikant sind. Dabei kamen z.B. sieben Motive nur in jeweils zwei Sequenzen vor, was 2,2 % des Datensatzes entspricht und somit um eine Größenordnung unter den von den Autoren genannten 20 % liegt. Nach Reduzierung des Datensatzes um die Membranproteine verblieben im Datensatz nur die, für die Adenylyl-Zyklasten charakteristischen Motive. Diese werden zwar beide nicht exprimiert, aufgrund der Zugehörigkeit zur gleichen Funktionsklasse und der Sequenzähnlichkeit (sieben gemeinsame Motive) kann allerdings nicht geschlossen werden, dass die Motive ursächlich für die Expressionshöhe sind.

MEME hat sich als empfindliches Instrument zur Ermittlung gemeinsamer Motive erwiesen. Die prinzipielle Limitation ist, dass Motive, die ursächlich für die Expressionshöhe sind, in mindestens zwei Sequenzen einer Kategorie vorhanden sein müssen um detektiert zu werden. Möglicherweise ermöglicht eine deutlich größere Anzahl von Sequenzen die Entdeckung von Motiven, die die Expression beeinflussen.

Bezüglich einer ungerichteten Suche nach Sequenzparametern kann gesagt werden, dass hierfür die Parameter eine gewisse Dominanz innerhalb des Datensatzes zeigen müssen, um entdeckt zu werden. Die Parameter, die eine Expression z.B. verhindern können, können sehr unterschiedlich beschaffen sein. Deshalb muss der Datensatz entsprechend groß sein um genug Sequenzen mit diesen Merkmalen zu enthalten. Mit dem analysierten Datensatz war über diesen Ansatz kein Merkmal zu finden, dass mit einer der Kategorien der Expressionshöhe assoziiert ist.

5.2.4 Die Verteilung sequenzbezogener Parameter

Im folgenden wird die Verteilung bestimmter Parameter auf die Kategorien der Expressionshöhe diskutiert.

Die Auswahl der analysierten Parameter erfolgte zunächst anhand der Literatur. Es wurden Parameter ausgewählt, für die ein Zusammenhang mit der Expressionshöhe von endogenen

Hefeproteinen sowie für Einzelfälle mit der Expressionshöhe heterolog exprimierter Proteine beschrieben worden ist, so z.B. „codon usage“ (Sharp und Cowe 1991; Brocca et al. 1998) oder AT-reiche Bereiche (Sinclair und Choy 2002). Darüber hinaus wurde die Verteilung von bekannten Parametern, die jedoch nicht mit der Expressionshöhe in Zusammenhang gebracht worden sind, wie z.B. isoelektrischer Punkt oder Ähnlichkeiten zu Hefeproteinen, analysiert.

5.2.4.1 Die Beurteilung der statistischen Signifikanz

Einleitend zur Betrachtung der Verteilung der analysierten Parameter auf die Kategorien der Expressionshöhe soll die Beurteilung der statistischen Signifikanz diskutiert werden.

Zunächst ist festzustellen, dass die vorliegenden Stichproben unabhängig sind, d.h. die Ausgangsmenge (in diesem Fall das humane Genom nach der Selektion durch das Teilprojekt Bioinformatik der PSF) ist groß im Vergleich zur Zahl der gezogenen Stichproben. Dies hat zur Folge, dass durch Auswahl einer Stichprobe (hier: cDNA) die Gesamtmenge so wenig reduziert wird, dass sich die Wahrscheinlichkeit der Ziehung einer bestimmten verbliebenen Stichprobe praktisch nicht ändert.

Zur Auswahl des richtigen Test zur Beurteilung von Signifikanz ist außerdem die Struktur der Daten zu betrachten. In den vorliegenden Fällen liegt aufgrund der Kategorisierung eine nichtstetige oder diskrete Verteilung vor. Die Kategorien sind nach Expressionshöhe sortierbar, so dass ordinalskalierte Werte vorliegen.

Weiterhin ist die Verteilung der Daten innerhalb der Kategorien zu berücksichtigen. Wie z.B. aus der Boxplot Darstellung der Verteilung seltener Codone (Abb. 19C) zu erkennen ist, kann weder von einer symmetrischen noch von einer Normalverteilung ausgegangen werden. Aus diesem Grund wurde in den Boxplots der Median und oberes und unteres Quartil dargestellt anstelle von arithmetischem Mittel und Standardabweichung. Zusätzlich erfordern andere Verteilungen als eine Normalverteilung verteilungsunabhängige Testverfahren. Als verteilungsunabhängiger Test zum gleichzeitigen Vergleich mehrerer ordinalskalierter und unabhängiger Stichproben wird der Kruskal-Wallis Test benutzt (z.B. Ariazi et al. 2002; Gentil Perret et al. 2002). Dieser Test prüft, ob sich mindestens eine der Kategorien von den anderen signifikant unterscheidet. Eine übliche Grenze für eine Beurteilung ist das 5 % Niveau, d.h. eine Wahrscheinlichkeit von 95 %, dass mindestens eine der Kategorien sich unterscheidet, wird bei der gegebenen Stichprobengröße als signifikant angesehen (z.B. Andrews et al. 2000; McCue et al. 2002; Chromiak et al. 2003).

Zur Beurteilung der Testresultate ist wichtig, dass verteilungsunabhängige Tests konservativer als parametrische sind, d.h. die Beurteilung von Signifikanz erfolgt vorsichtiger in verteilungsunabhängigen als in parametrischen Tests (zusammengefasst von Sachs 2002).

5.2.4.2 Die Verteilung Nukleotidsequenz basierte Parameter

Sowohl für *S. cerevisiae* als auch für *P. pastoris* wurde frühzeitige Transkriptionstermination an AT-reichen Bereichen als kritisch für eine heterologe Expression beschrieben (Romanos et al. 1991; Scorer et al. 1993; Milek et al. 2000; Gurkan und Ellar 2003). Dieses Phänomen kommt dadurch zustande, dass Terminationssignale in Hefe deutlich weniger definiert sind als z.B. in Säugern (Zhao et al. 1999, siehe auch Einleitung und Abb. 1). Bei der Transkription von Säugersequenzen in Hefen können AT-reiche Bereiche, die im Säugersystem nicht relevant sind, als Transkriptionsterminationssignal wirken. Eine eindeutige Identifizierung von Hefeterminationssignalen ist wegen ihrer undefinierten Natur häufig nicht möglich (Romanos et al. 1991).

Der in dieser Arbeit gefundene Zusammenhang zwischen einem niedrigen Gehalt an solchen Bereichen und einer guten Expression zeigt, dass AT-reiche Bereiche nicht nur in Einzelfällen ein wichtiger Parameter sind, sondern ein niedriger Gehalt eine generelle Voraussetzung für eine gute Expression in *P. pastoris* ist.

Terminationssignale in Hefe bestehen aus vier Elementen (siehe Abb. 1). Ein schwacher Konsensus eines dieser Elemente kann von einem starken Konsensus der anderen Elemente kompensiert werden. Es kann auch zu „schwachen“ Elementen kommen, die zu einer geringeren Terminationseffizienz führen (Graber et al. 1999). Die Erkenntnis, dass ein niedriger Gehalt an AT-reichen Bereichen eine Voraussetzung für eine gute Expression zu sein scheint, aber Sequenzen mit deutlich höheren Werten noch zu detektierbarem Protein führen, könnte auf solche „schwachen“ Terminationssignale zurückzuführen sein. Weiterführende Experimente zu mRNA Gehalt bzw. Länge der entstehenden Transkripte könnten hier Aufschluss geben.

In jüngerer Zeit wurden zunehmend theoretische Ansätze zur Vorhersage von Terminationssignalen in Hefe entwickelt (Graber et al. 2002). Wenn diese Algorithmen in der Lage sein werden, auch „schwache“ Terminationssignale zuverlässig zu erkennen, würde das ebenfalls eine Möglichkeit der genaueren Analyse bieten.

Sequenzen mit einem geringen Gehalt an AT-reichen Regionen werden nicht notwendigerweise gut exprimiert. Dies zeigt, dass es nicht eine einzelne Limitation bei heterologer Expression gibt.

Zu den Verteilungen von GC-Gehalt und GC3s ist anzumerken, dass beide Parameter positiv mit der Transkriptionshöhe von 4977 offenen Leserahmen in *S. cerevisiae* korrelieren (Marin et al. 2003). Der Parameter GC3s ist für eine Beurteilung des Einflusses der Nukleotidkomposition deshalb von Interesse, da dieser Wert im Gegensatz zum GC-Gehalt nicht von der Aminosäurezusammensetzung des Proteins abhängt und deshalb keinem Selektionsdruck ausgehend von der Funktionalität des kodierten Proteins unterliegt.

Die Verteilung der Daten zeigt, dass für beide Parameter kein Zusammenhang zur Expressionshöhe besteht. Möglicherweise ist der Selektionsdruck, der zu dem von Marin und Mitarbeitern (2003) beschriebenen Zusammenhang geführt hat, zu schwach, um eine heterologe Expression zu beeinflussen.

Die „codon usage“ als Einflussgröße auf die Expressionshöhe wurde betrachtet, da in *S. cerevisiae* schon relativ früh Präferenzen für bestimmte Codone in hochexprimierten Genen gefunden wurden (Sharp et al. 1986). Eine positive Korrelation zwischen der Häufigkeit bestimmter synonyme Codone und der entsprechenden tRNA wurde für mehrere Organismen gezeigt (Ikemura 1982; Ikemura 1985; Sharp et al. 1986). Ein Modell der Koevolution von „codon usage“ und tRNA Menge erklärt diese Tendenz über Vorteile bei der Translation durch besser zur Verfügung stehende Aminoacyl-tRNAs (Bulmer 1987). Dieser relativ einfache Zusammenhang wurde allerdings in jüngerer Zeit in Frage gestellt. So gibt es in verschiedenen Eukaryonten – unter anderem in *S. cerevisiae* – eine negative Korrelation zwischen der Frequenz optimaler Codone und der Proteinlänge (Moriyama und Powell 1998; Comeron et al. 1999; Duret und Mouchiroud 1999). Dieser Zusammenhang ist nicht durch eine geringere Länge hochexprimierter Proteine bedingt und zumindest für *Drosophila melanogaster* in etwa genauso stark ausgeprägt wie der Zusammenhang zwischen „codon usage“ und Expressionshöhe (Duret und Mouchiroud 1999). In hochexprimierten *D. melanogaster* Alkoholdehydrogenase Genen ist das Potential zur mRNA Sekundärstrukturbildung höher als in schwach exprimierten Genen (Carlini et al. 2001). Gygi und Mitarbeiter (1999) fanden bei der Analyse der Expression von 150 *S. cerevisiae* Genen, dass die „codon usage“ keine Vorhersage weder von der Transkriptionshöhe noch von der Proteinmenge erlaubt. Alle diese Erkenntnisse führten zu dem Schluss, dass über die Verfügbarkeit von beladenen tRNAs hinaus weitere Selektionsmechanismen die Auswahl synonyme Codone beeinflussen (Duret 2002). Einen solchen Mechanismus schlagen Elf und Mitarbeiter (2003) vor. Die Autoren fanden eine Überrepräsentation von seltenen Codonen in Genen, die für Enzyme in Aminosäuresynthesewegen kodieren. Das abgeleitete Modell sieht seltene Codone bzw. die entsprechenden Aminoacyl-tRNAs als Reserve für den Fall von

Aminosäuremangel. In diesem Fall würden die seltenen tRNAs beladen vorliegen, da die für bevorzugte Codone vermehrt „verbraucht“ werden. Dies ermöglicht der Zelle eine Synthese der benötigten Syntheseeenzyme.

Erkenntnisse, die für einen untergeordneten Einfluss der „codon usage“ auf die Expressionshöhe sprechen, gibt es auch für heterologe Expression in *P. pastoris*. Ein zunächst hinsichtlich der „codon usage“ optimiertes Gen für humane Glucocerebrosidase wird um den Faktor 10,6 besser exprimiert als das humane Allel. Durch die Anpassung der „codon usage“ steigt jedoch der GC-Gehalt; ein Kontrollkonstrukt mit erhöhtem GC-Gehalt, jedoch humaner „codon adaptation“ wird um den Faktor 7,5 höher exprimiert. Das deutet auf den GC-Gehalt als Haupteinflussgröße hin (Sinclair und Choy 2002).

Ein weiterer Gesichtspunkt ist die Variation der „codon adaptation“ im *S. cerevisiae* Genom. Eine Analyse von 6217 offenen Leserahmen in *S. cerevisiae* ergab, dass der Durchschnitt des CAI 0,107 bei einer Standardabweichung von 0,017 beträgt (Coghlan und Wolfe 2000). Der größte Teil der *S. cerevisiae* Gene zeigt also eine „codon adaptation“, die gut im Bereich der in dieser Arbeit klonierten Sequenzen liegt (die Durchschnittswerte der Kategorien liegen zwischen 0,090 und 0,100). Eine „codon adaptation“ in dieser Größenordnung kann also nicht zu einer kompletten Inhibierung der Expression in Hefe führen.

Negative Auswirkungen auf eine sehr gute Expression wären denkbar, da eine hohe „codon adaptation“ auch in *S. cerevisiae* nur bei hochexprimierten Genen vorliegt (Sharp et al. 1986), ist jedoch aus den hier vorliegenden Daten nicht abzuleiten. Wenn eine Steigerung der Expression in *P. pastoris* durch Anpassung der „codon usage“ erreicht wurde, wurde diese auf einen CAI von 1,0 (Sinclair und Choy 2002) gesteigert. Eine abgestufte Steigerung der „codon adaptation“ wurde nicht untersucht. Sequenzen mit einem CAI dieser Größenordnung sind in dieser Arbeit nicht enthalten. Möglicherweise ist ein positiver Einfluss auf die Expression erst bei sehr guter Anpassung der „codon usage“ zu sehen.

Die Anzahl der effektiven Codone (N_c) und die „codon adaptation“ an *S. cerevisiae* wurden ermittelt, da diese beiden Indices eine geringe Empfindlichkeit gegen unterschiedliche Längen der untersuchten Sequenzen zeigen (Comeron und Aguade 1998). Das eignet diese Indices für den Vergleich der hier vorliegenden, verschieden langen Sequenzen. Allerdings zeigten Simulationen, dass der N_c bei kurzen Sequenzen (300 – 500 Bp) zu einer leichten Unterschätzung tendiert (Moriyama und Powell 1998). Aufgrund der Selektion der Sequenzen durch die PSF sind viele solcher Sequenzen vertreten; bei der Interpretation ist dies zu berücksichtigen.

Da die bevorzugten Codone in *P. pastoris* ähnlich denen in *S. cerevisiae* sind (Sinclair und Choy 2002), ist eine Messung der „codon adaptation“ gegen *S. cerevisiae* zulässig.

Die vorliegende Arbeit zeigt, dass eine Variation der „codon usage“ in dem hier untersuchten Bereich keinen generellen Einfluss auf die Expressionshöhe hat. Das impliziert, dass die Verfügbarkeit bestimmter beladener tRNAs keine Limitation für die Translation darstellt. Dies ist in Übereinstimmung mit den genannten Hypothesen für endogene *S. cerevisiae* Gene, die die „codon adaptation“ als Determinante für eine starke Expression in Frage stellen.

Der Einfluss seltener Codone auf die Expressionshöhe in Hefe wurde weit weniger detailliert untersucht als der der „codon adaptation“ der Gesamtsequenz. Herrick und Mitarbeiter (1990) fanden bei der Analyse von 20 *S. cerevisiae* Genen eine Häufung von seltenen Codonen in instabilen Transkripten. Ein Einfluss von seltenen Codonen auf heterologe Expression in Hefe ist bisher nicht untersucht worden. In dieser Arbeit kann kein Einfluss von seltenen Codonen auf die heterologe Expression in *P. pastoris* festgestellt werden. Sollten seltene Codone allerdings keinen Translationsnachteil mit sich bringen, sondern der Vorratshaltung im Falle von Aminosäuremangel dienen (Elf et al. 2003, siehe oben), wäre dies auch nicht zu erwarten. Yelin und Schuldiner (2001) fanden einen Einfluss der „codon usage“ am 5'-Ende der kodierenden Sequenz auf die Expression eines Monoamintransporters aus Ratte in *S. cerevisiae*. Durch den N-terminalen 6*His-tag sowie die nachfolgende Restriktionsschnittstelle im Vektor pPICH5 sind in unserem Falle die ersten zehn Codone für alle Sequenzen identisch. Aus diesem Grunde kann hier ein solcher Vergleich nicht stattfinden.

5.2.4.3 Die Verteilung Proteinsequenz basierter Parameter

Interessanterweise ist ein hoher isoelektrischer Punkt assoziiert mit keiner detektierbaren Expression. Schwartz und Mitarbeiter (2001) fanden bei Datenbankanalysen, dass bei Eukaryonten, Archaeobakterien und Prokaryonten die isoelektrischen Punkte von Proteinen mit deren Lokalisation zusammenhängen. Membranproteine haben pI-Werte um neun, cytosolische Proteine um fünf. Die hohen pI-Werte von Membranproteinen könnten mit der negativ geladenen Oberfläche von Biomembranen zusammenhängen. Basische Aminosäuren, zumindest in an der Membranaußenseite lokalisierten Bereichen von Proteinen könnten durch elektrostatische Wechselwirkung die Lokalisation der Proteine stabilisieren. Es gibt keine Hypothese, warum cytosolische Proteine bevorzugt einen pI von unter sieben zeigen (Schwartz et al. 2001). Die Tatsache, dass Proteine mit einem hohen pI sich nicht cytosolisch

exprimieren lassen, könnte auf einen regulativen Mechanismus in Hefezellen hindeuten, der solche Proteine z.B. degradiert.

Die Löslichkeit von Proteinen ist bei pH-Werten nahe dem pI am geringsten (Arakawa und Timasheff 1985), somit könnte der Mangel an Proteinen mit einem pI um sieben durch den meistens neutralen pH-Wert im Cytoplasma (Melvin und Shanks 1996; Carmelo et al. 1997) bedingt sein. In Hefe kann der cytosolische pH allerdings auf bis zu 5,5 während der stationären Phase absinken (Imai und Ohno 1995). Das entspricht dem pI der meisten cytoplasmatischen Proteine.

Zusätzlich bemerkenswert ist die Verteilung der pI-Werte innerhalb der Kategorie der nicht exprimierten Proteine (siehe Abb. 20D). Man sieht eine bimodale Verteilung, d.h. es gibt viele Proteine mit einem pI um sechs, nur wenige mit neutralem pI und eine zweite Häufung mit einem pI um neun. Der Mangel an Proteinen mit pI um sieben könnte durch den ubiquitären Mangel an derartigen cytosolischen Proteinen (Schwartz et al. 2001) bedingt sein. Unter der Voraussetzung, dass tatsächlich der pI die Ursache ist, kann hier wieder abgeleitet werden, dass es keinen Parameter gibt, der allein für die Expressionshöhe verantwortlich gemacht werden kann. Es gibt mehrere Sequenzen, die sich trotz niedrigem pIs nicht exprimieren lassen.

Zu den Werten für den isoelektrischen Punkt ist zu sagen, dass diese notwendigerweise unexakt sind, da in die Berechnung keine Information über mögliche Einflüsse der Faltung auf ionisierbare Gruppen eingeht. Es wurde jedoch gezeigt, dass theoretische pI-Werte relativ gut mit gemessenen übereinstimmen (Sillero und Ribeiro 1989).

Zum Zusammenhang zwischen der Länge eines Proteins und dessen Expressionshöhe gibt es für *S. cerevisiae* mehrere Studien. Es wurde gezeigt, dass es keinen Zusammenhang zwischen mRNA Stabilität und Länge gibt (Herrick et al. 1990). Ebenso ist keine Korrelation zwischen Genlänge und Transkriptionshöhe zu sehen (Jansen und Gerstein 2000). Es gibt eine negative Korrelation zwischen Protein-Länge und „codon adaptation“, die möglicherweise durch eine Kompensation des höheren Synthesaufwandes für längere Proteine durch Benutzung bevorzugter Codone zustande kommt (Moriyama und Powell 1998).

In dieser Arbeit ist ebenfalls kein Zusammenhang zwischen Proteinlänge und Expressionshöhe zu sehen. Hier ist zu berücksichtigen, dass durch die Vorselektion in der PSF keine Proteine länger als 500 Aminosäuren bearbeitet wurden. Innerhalb dieses Bereiches ist die Größe für *P. pastoris* möglicherweise kein kritischer Faktor. Selbst wenn es für endogene Proteine eine Korrelation aufgrund einer Selektion kurzer Proteine wegen energetischer Vorteile gibt, muss diese sich nicht in einem Expressionsexperiment

wiederfinden. In einem solchen Fall hat keine Selektion stattgefunden, die kürzere codierende Sequenzen bei gleicher Funktionalität des Proteins bevorzugt.

Weiterhin wurde untersucht, ob sich Zusammenhänge zwischen einer Ähnlichkeit der Proteine zu Hefeproteinen und ihrer Exprimierbarkeit ergeben. Es wurden hier zwei verschiedene Ansätze gewählt. Zunächst wurde mittels BLAST die Ähnlichkeit zu Hefesequenzen quantifiziert. Der BLAST-Algorithmus findet lokale Ähnlichkeiten in verglichenen Sequenzen. Je größer der jeweilige gemeinsame Bereich, desto höher der für diesen Bereich vergebene „similarity score“. Kommen in zwei verglichenen Sequenzen mehrere, voneinander getrennte ähnliche Bereiche vor, so gibt es mehrere Hits (Mount 2001). Aus diesem Grund wurde nicht nur die Verteilung der jeweils besten Hits, sondern auch die Summe aller Hits einer Sequenz betrachtet. Damit wird berücksichtigt, dass nur einmal vorkommende Ähnlichkeiten mit einem hohen „similarity score“ nicht höher bewertet werden als viele über die Sequenz verteilte mit jeweils niedrigen „similarity scores“.

Die Analyse mittels SUPERFAMILY basiert auf einer Einordnungen der analysierten Proteine aufgrund von Gemeinsamkeiten mit bestimmten Strukturfamilien im Gegensatz zu den rein sequenzorientierten Vergleichen durch BLAST (Gough et al. 2001).

Mit beiden Herangehensweisen konnte kein signifikanter Zusammenhang zu einer der Kategorien der Expressionshöhe festgestellt werden. Wenn die Ähnlichkeit zu Hefeproteinen einen Parameter für eine erfolgreiche Expression darstellt, würde das voraussetzen, dass Hefe einen Erkennungsmechanismus für eigene bzw. fremde Proteine besitzt. Es ist schwer vorstellbar, dass in Hefe eine bis heute unentdeckte Erkennung und Selektion von Proteinen aufgrund der Primärstruktur – vergleichbar dem Immunsystem höherer Organismen – existiert. Selektionsmechanismen basierend auf Sekundär- oder Tertiärstruktur wären hingegen vorstellbar. So ist bekannt, dass *S. cerevisiae* Zellen eine Akkumulation von ungefalteten Proteinen im Endoplasmatischen Retikulum (ER) detektieren und darauf reagieren können (Welihinda et al. 1999). Interessanterweise liegt bei beiden Darstellungen der strukturellen Ähnlichkeit der Proteine zu *S. cerevisiae* Proteinen der Median der Kategorie der gut exprimierten Proteine deutlich über denen der anderen Kategorien. Dieser Unterschied ist allerdings nicht auf dem 5 % Niveau signifikant. Eine Analyse mit einer größeren Anzahl von Sequenzen, insbesondere in der Kategorie der guten Exprimierer wäre interessant.

Es kann ausserdem kein Zusammenhang zwischen der Verteilung von Sekundärstrukturelementen und der Expressionshöhe festgestellt werden. Es muss berücksichtigt werden, dass Sekundärstrukturvorhersagen fehlerbehaftet sind, so ordnete

PSI-PRED in einer Evaluierung verschiedener Vorhersagealgorithmen im Durchschnitt 77,3 % aller Aminosäurereste dem korrekten Sekundärstrukturelement zu (Orengo et al. 1999). Mit einem entsprechenden Fehler muss auch hier gerechnet werden.

Eine Diskrimination von Proteinen aufgrund von Sekundärstrukturen würde einen entsprechenden Erkennungsapparat voraussetzen. Sekundärstrukturen alleine sind möglicherweise ein zu unspezifisches Merkmal um – analog zur Erkennung ungefalteter Proteine im ER – in Hefe erkannt zu werden.

5.2.4.4 Proteindegradationssignale

Weiterhin wurden verschiedene Parameter, die einen Einfluss auf die Proteinstabilität *in vivo* haben können, betrachtet.

Zunächst sind hier PEST-Motive zu nennen. Diese hydrophile Motiv ist in vielen Proteinen mit geringer Halbwertszeit zu finden und kann Instabilität auf Proteine in verschiedenen Spezies übertragen. Der Abbau der Proteine erfolgt wahrscheinlich über das 26S-Proteasom (zusammengefasst von Rechsteiner und Rogers 1996).

Es ist für alle gewählten Formen der Quantifizierung kein Zusammenhang zwischen Vorkommen von PEST-Motiven und Expressionshöhe zu sehen. Allerdings beinhalten nur wenige Sequenzen im Datensatz tatsächlich Motive, deren Ähnlichkeit zum Konsensus hoch genug ist, um von echten PEST-Motiven zu sprechen (PEST „similarity score“ > 50) (Rechsteiner und Rogers 1996). Diese kommen jedoch in allen vier Kategorien vor. Die Verteilung der übrigen Motive ist möglicherweise gar nicht relevant, da diese nicht als PEST-Motive wirksam sind.

Ein additiver Effekt schwacher Motive ist ebenfalls nicht nachzuweisen, da auch die Summe aller gefundenen „similarity scores“ pro Protein keinen Zusammenhang zur Expressionshöhe zeigt.

Hydrophobe Bereiche – im Gegensatz zu der oben betrachteten Gesamthydrophobizität des Proteins - wurden aus zwei Gründen im Zusammenhang mit Proteindegradation betrachtet. Zum Einen können sie als Signale für eine Ubiquitinierung und folgende Degradation des Proteins fungieren (Gilon et al. 2000), zum Anderen können hydrophobe Bereiche, die während der Proteinsynthese nicht im Inneren des Moleküls verborgen liegen, zu einer Aggregation führen (Dobson und Karplus 1999). Aggregierte Proteine haben in *S. cerevisiae* eine geringere Halbwertszeit als nativ gefaltete (Saris et al. 1997).

In dieser Arbeit kann kein Zusammenhang zwischen hydrophoben Bereichen der Proteine und der Expressionshöhe festgestellt werden. Dies kann darin begründet sein, dass das

Vorkommen von unlöslichen rekombinanten Proteinen in Hefe nur in Einzelfällen berichtet wurde (z.B. Weik et al. 1998). In *S. cerevisiae* sind mehrere Chaperone bekannt, die an die aus dem Ribosom austretende Polypeptidkette binden und so eine Aggregation verhindern (Craig et al. 2003), das ist möglicherweise der Grund, warum eine Aggregation während der heterologen Expression in Hefe keine ausreichend große Rolle spielt, um hier beobachtet werden zu können.

Lysinreste wurden als potentielle Stellen für eine Ubiquitinierung und folgender Degradation (Hochstrasser 1996) in Betracht gezogen. Ein Zusammenhang zu dem Vorkommen von Lysinresten ist allerdings nicht erkennbar, auch dann nicht, wenn man die Oberflächenwahrscheinlichkeit des jeweiligen Restes berücksichtigt. Dazu ist zu sagen, dass nicht alle Lysinreste ubiquitiniert werden. Zusätzliche Signale – wie z.B. PEST-Motive oder Phosphorylierungen (Craig und Tyers 1999) – spielen eine Rolle, um den Ubiquitinierungsapparat der Zelle zu steuern. Insofern müssen Lysine nicht zu einem Abbau führen und aus der hier vorliegenden Verteilung kann nicht geschlossen werden, dass Proteindegradation über Ubiquitinierung keine Einflussgröße auf eine heterologe Expression ist.

Neben der ubiquitingesteuerten Degradation sind auch ubiquitinunabhängige Mechanismen bekannt (Hoyt et al. 2003). Zusätzlich zu einer Ubiquitinierung können also auch andere Mechanismen hier eine Rolle spielen. Die Schlussfolgerung, dass Proteindegradation bei heterologer Expression in *P. pastoris* keine deutliche Rolle spielt, ist also nicht zu ziehen.

5.2.4.5 Die Kombination verschiedener Parameter

Es ist bekannt, dass es bei heterologer Expression in *P. pastoris* zur Überlagerung mehrerer Effekte kommen kann (Sinclair und Choy 2002). Dies kann eine statistisch signifikante Zuordnung eines der Parameter zu einer bestimmten Kategorie erschweren bzw. die Relevanz eines Parameters maskieren. Deshalb wurden kombinatorische Darstellungen von zwei Parametern vorgenommen. Dies könnte zeigen, ob eine bestimmte Merkmalskombination ein Ausschlusskriterium für eine Expression darstellt. Im Falle der Kombination von AT-reichen Bereichen und der Anzahl der Lysinreste auf der Proteinoberfläche (Abb. 27A) kann ein solcher Effekt vermutet werden. Eine Signifikanz kann allerdings nicht nachgewiesen werden. Es ist nicht festzustellen, ob Ausreißer in einer der Kategorien (siehe z.B. einige Sequenzen mit einer hohen Anzahl von Lysinresten in der Kategorie der nicht exprimierten Proteine, Abb. 22A – D) zufällig sind oder ein Ausschlusskriterium für eine Expression dieser Proteine

darstellen. Auch hier könnte die Analyse einer höheren Anzahl von Sequenzen eine Signifikanz zeigen.

5.3 Fazit und Ausblick

Es wurde eine Parallelisierung der Expressionskontrolle erreicht. Durch die Reduzierung des Kultivierungsmaßstabes auf 2 mL und die chemische Lyse der Zellen anstelle einer mechanischen, sind die limitierenden Schritte des Prozesses neben den Klonierungen die Transformation der Zellen und Analyse der Lysate. Eine mögliche Verbesserung liegt in einer Vereinfachung der Klonierungen durch die Verwendung Ligations-unabhängiger Klonierungssysteme (Hammarstroem et al. 2002). Zur Parallelisierung der Transformation von *P. pastoris* kann z.B. die Effizienz von chemischen Transformationsmethoden erhöht werden. Bei ausreichender Effizienz kann im 96'er Maßstab transformiert werden (Holz et al. 2003).

Der zweite Teil der Arbeit belegt, dass durch den gewählten komparativen Ansatz Parameter, die die Expression beeinflussen, identifiziert werden können. Es konnte zum Einen gezeigt werden, dass der Einfluss eines zuvor für Einzelfälle nachteilig beschriebenen Merkmals (das Vorkommen AT-reicher Bereiche) ein allgemeines Kriterium für die Exprimierbarkeit darstellt. Zum Anderen konnte mit dem isoelektrischen Punkt ein Parameter identifiziert werden, der zuvor noch nicht als relevant für heterologe Expression erkannt wurde. Die „codon adaptation“, ein in diesem Zusammenhang viel diskutierter Parameter, hat zumindest in den hier untersuchten Bereichen keinen allgemeinen Einfluss auf die Expressionshöhe.

Diese Ergebnisse zeigen Möglichkeiten sowohl für eine gezielte Optimierung von Sequenzen zur Expression in *P. pastoris* als auch zur Auswahl eines geeigneten Expressionswirtes auf. So muss ein hoher Gehalt an AT-reichen Bereichen vermieden werden, um in *P. pastoris* zu einer guten Expression zu kommen. Eine Verbesserung der „codon adaptation“ kann innerhalb der hier untersuchten Grenzen vernachlässigt werden. Eine Verschiebung des isoelektrischen Punktes ist notwendigerweise mit einer Veränderung des Proteins verbunden. Ist dies nicht sinnvoll, so können für entsprechende Proteine Expressionsexperimente in *P. pastoris* vermieden und alternative Wirte in Betracht gezogen werden.

Als ein weiterer theoretisch zu analysierender Parameter wäre eine Untersuchung auf globale Strukturmerkmale, sogenannte „folds“ zu nennen. Eine Analyse des *S. cerevisiae* Transkriptoms ergab eine Anreicherung an Sequenzen, die für α - β Mixstrukturen kodieren (Jansen und Gerstein 2000).

Eine experimentelle Analyse bzw. Überprüfung der gefundenen Parameter wäre grundsätzlich interessant. So müsste eine vorzeitige Terminierung der Transkription von Sequenzen, die AT-reiche Bereiche enthalten, experimentell zu zeigen sein. Unterschiede in den mRNA-Stabilitäten könnten mit anderen Ansätzen erfasst werden. Zum einen könnte unter Nutzung der bestehenden Sammlung an Klonen versucht werden, die Transkription mittels Thiolutin zu hemmen, um so kürzere Zeiten experimentell erfassen zu können. Eine weitere Möglichkeit besteht in einer generellen Stabilisierung der Transkripte durch die Verwendung eines alternativen Terminators, der zu einer stabilisierenden 3'-UTR führt. Dies würde nicht nur zu einer Vereinfachung der Halbwertszeitmessung führen, sondern möglicherweise auch zu generell höheren Proteinausbeuten führen.

Aufgrund der oben erwähnten Ausreißer-Problematik gibt es weitere Parameter, die eine weitergehende Analyse interessant machen. Interessant wäre hier eine gezielte Mutagenese von Sequenzen, die für bestimmte Parameter Extremwerte aufweisen. Das könnte Aufschluss darüber geben, ob die entsprechenden Parameter in extremer Ausprägung einen Einfluss auf die Expression haben.

6 Zusammenfassung

Im ersten Teil der vorliegenden Arbeit wurde ein System entwickelt und angewandt, um *Pichia pastoris* Expressionsklone für humane cDNAs im Hochdurchsatzverfahren hinsichtlich ihrer Expression zu überprüfen und zu klassifizieren (Boettner et al. 2002). Im zweiten Teil wurden die mit diesem Verfahren erstellten Expressionsklone auf sequenzbasierte Parameter untersucht, die mit der Expressionshöhe korrelieren.

Die entwickelte Methode erlaubt die Anzucht und Induktion der Proteinexpression im 2 mL Maßstab. Der Aufschluss der Zellen erfolgt chemisch in SDS-PAGE Probenpuffer, so dass das Lysat direkt in eine Gelelektrophorese eingesetzt und mittels Western-Blot auf das Vorhandensein des rekombinanten Proteins überprüft werden kann.

Die so erstellten Klone wurden entsprechend der beobachteten Expressionshöhe in vier Kategorien eingeteilt.

Ungerichtete Untersuchungen auf eine Sortierung der Sequenzen anhand von nicht vordefinierten Sequenzmerkmalen in phylogenetischen Bäumen oder gemeinsamen Motiven ergaben keine Merkmale, die sich bestimmten Kategorien der Expressionshöhe zuordnen lassen.

Die Analyse der Verteilung bestimmter Sequenzmerkmale auf die Kategorien der Expressionshöhe ergab einen signifikanten Zusammenhang zwischen einer guten Expression eines Proteins und einem geringen Gehalt an AT-reichen Bereichen in der codierenden Sequenz. Ein hoher Gehalt führt möglicherweise zu einer frühzeitigen Termination der Transkription durch Hefe an solchen AT-reichen Bereichen. Eine genaue Zuordnung von Hefe-Terminationssignalen zu den Sequenzen ist aufgrund der variablen Natur dieser Signale in Hefe nicht möglich.

Ein weiterer Zusammenhang wurde zwischen einer nicht detektierbaren Expression und einem hohen isoelektrischen Punkt des Proteins beobachtet. Im Zusammenhang mit dem Phänomen, dass sowohl in Pro- als auch in Eukaryonten cytosolische Proteine einen relativ niedrigen pI aufweisen, könnte das ein Hinweis auf einen regulatorischen Mechanismus sein.

Es konnte gezeigt werden, dass durch einen vergleichenden Ansatz Parameter identifiziert werden können, die signifikant mit der Expressionshöhe von Proteinen zusammenhängen.

Die Ergebnisse ermöglichen Schlüsse für eine Veränderung von Sequenzen für eine bessere Exprimierbarkeit bzw. eine Vorauswahl von Proteinen zur Expression in *P. pastoris*.

7 Literatur

- Abbott, A. (2000) Structures by numbers. *Nature* **408**: 130-2.
- Acosta-Rivero, N., Sanchez, J. C. und Morales, J. (2002) Improvement of human interferon HUIFNalpha2 and HCV core protein expression levels in *Escherichia coli* but not of HUIFNalpha8 by using the tRNA(AGA/AGG). *Biochem Biophys Res Commun* **296**: 1303-9.
- Alexeyev, M. F. und Winkler, H. H. (1999) Gene synthesis, bacterial expression and purification of the *Rickettsia prowazekii* ATP/ADP translocase. *Biochim Biophys Acta* **1419**: 299-306.
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W. und Lipman, D. J. (1990) Basic local alignment search tool. *J Mol Biol* **215**: 403-10.
- Andrews, J., Bouffard, G. G., Cheadle, C., Lu, J., Becker, K. G. und Oliver, B. (2000) Gene discovery using computational and microarray analysis of transcription in the *Drosophila melanogaster* testis. *Genome Res* **10**: 2030-43.
- Andrin, C., Corbett, E. F., Johnson, S., Dabrowska, M., Campbell, I. D., Eggleton, P., Opas, M. und Michalak, M. (2000) Expression and purification of mammalian calreticulin in *Pichia pastoris*. *Protein Express Purif* **20**: 207-15.
- Arakawa, T. und Timasheff, S. N. (1985) Theory of protein solubility. *Methods Enzymol* **114**: 49-77.
- Ariazi, E. A., Clark, G. M. und Mertz, J. E. (2002) Estrogen-related receptor α and estrogen-related receptor γ associate with unfavourable and favourable biomarkers, respectively, in human breast cancer. *Cancer Res* **62**: 6510-8.
- Bailey, L. B. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*: 28 - 36.
- Baim, S. B. und Sherman, F. (1988) mRNA structures influencing translation in the yeast *Saccharomyces cerevisiae*. *Mol Cell Biol* **8**: 1591-601.
- Balbas, P. (2001) Understanding the art of producing protein and nonprotein molecules in *Escherichia coli*. *Mol Biotechnol* **19**: 251-67.
- Batard, Y., Hehn, A., Nedelkina, S., Schalk, M., Pallett, K., Schaller, H. und Werck-Reichhart, D. (2000) Increasing expression of P450 and P450-reductase proteins from monocots in heterologous systems. *Arch Biochem Biophys* **379**: 161-9.
- Bateman, A., Birney, E., Durbin, R., Eddy, S. R., Howe, K. L. und Sonnhammer, E. L. (2000) The Pfam protein families database. *Nucleic Acids Res* **28**: 263-6.
- Bennetzen, J. L. und Hall, B. D. (1982) Codon selection in yeast. *J Biol Chem* **257**: 3026-31.
- Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. und Wheeler, D. L. (2003) GenBank. *Nucleic Acids Res* **31**: 23-7.
- Benson, D. A., Karsch-Mizrachi, I., Lipmann, D. J., Ostell, J. und Wheeler, D. L. (2003) GenBank. *Nucleic Acids Res* **13**: 23-7.
- Berrin, J. G., Williamson, G., Puigserver, A., Chaix, J. C., McLauchlan, W. R. und Juge, N. (2000) High-level production of recombinant fungal endo- β -1,4-xylanase in the methylotrophic yeast *Pichia pastoris*. *Prot Expr Purif* **19**: 179-87.

- Bettany, A. J., Moore, P. A., Cafferkey, R., Bell, L. D., Goodey, A. R., Carter, B. L. und Brown, A. J. (1989) 5'-secondary structure formation, in contrast to a short string of non-preferred codons, inhibits the translation of the pyruvate kinase mRNA in yeast. *Yeast* **5**: 187-98.
- Bisht, H., Chugh, D. A., Swaminathan, S. und Khanna, N. (2001) Expression and purification of dengue virus type 2 envelope protein as a fusion with hepatitis B surface antigen in *Pichia pastoris*. *Protein Express Purif* **23**: 84-96.
- Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M. C., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S. und Schneider, M. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res* **31**: 365-70.
- Boettner, M., Prinz, B., Holz, C., Stahl, U. und Lang, C. (2002) High-throughput screening for expression of heterologous proteins in the yeast *Pichia pastoris*. *J Biotechnol* **99**: 51-62.
- Brierley, R. A., Bussineau, C., Kosson, R., Melton, A. und Siegel, R. S. (1990) Fermentation development of recombinant *Pichia pastoris* expressing the heterologous gene: bovine lysozyme. *Ann N Y Acad Sci* **589**: 350-62.
- Brocca, S., Schmidt-Dannert, C., Lotti, M., Alberghina, L. und Schmid, R. D. (1998) Design, total synthesis, and functional overexpression of the *Candida rugosa* lip1 gene coding for a major industrial lipase. *Protein Sci* **7**: 1415-22.
- Bucheler, U. S., Werner, D. und Schirmer, R. H. (1990) Random silent mutagenesis in the initial triplets of the coding region: a technique for adapting human glutathione reductase-encoding cDNA to expression in *Escherichia coli*. *Gene* **96**: 271-6.
- Bucheler, U. S., Werner, D. und Schirmer, R. H. (1992) Generating compatible translation initiation regions for heterologous gene expression in *Escherichia coli* by exhaustive perShine-Dalgarno mutagenesis. Human glutathione reductase cDNA as a model. *Nucleic Acids Res* **20**: 3127-33.
- Bulmer, M. (1987) Coevolution of codon usage and transfer RNA abundance. *Nature* **325**: 728-30.
- Butz, J. A., Niebauer, R. T. und Robinson, A. S. (2003) Co-expression of molecular chaperones does not improve the heterologous expression of mammalian G-protein coupled receptor expression in yeast. *Biotechnol Bioeng* **84**: 292-304.
- Calderone, T. L., Stevens, R. D. und Oas, T. G. (1996) High-level misincorporation of lysine for arginine at AGA codons in a fusion protein expressed in *Escherichia coli*. *J Mol Biol* **262**: 407-12.
- Caponigro, G., Muhlrads, D. und Parker, R. (1993) A small segment of the MAT alpha 1 transcript promotes mRNA decay in *Saccharomyces cerevisiae*: a stimulatory role for rare codons. *Mol Cell Biol* **13**: 5141-8.
- Carlini, D. B., Chen, Y. und Stephan, W. (2001) The relationship between third-codon position nucleotide content, codon bias, mRNA secondary structure and gene expression in the drosophilid alcohol dehydrogenase genes *Adh* and *Adhr*. *Genetics* **159**: 623-33.
- Carmelo, V., Santos, H. und Sa-Correia, I. (1997) Effect of extracellular acidification on the activity of plasma membrane ATPase and on the cytosolic and vacuolar pH of *Saccharomyces cerevisiae*. *Biochim Biophys Acta* **1325**: 63-70.
- Cereghino, J. L. und Cregg, J. M. (2000) Heterologous protein expression in the methylotrophic yeast *Pichia pastoris*. *FEMS Microbiol Rev* **24**: 45-66.

- Chen, C. Y. und Shyu, A. B. (1995) AU-rich elements: characterization and importance in mRNA degradation. *Trends Biochem Sci* **20**: 465-70.
- Chen, Y., Jin, M., Egborge, T., Coppola, G., Andre, J. und Calhoun, D. H. (2000) Expression and characterization of glycosylated and catalytically active recombinant human alpha-galactosidase A produced in *Pichia pastoris*. *Protein Expr Purif* **20**: 472-84.
- Christendat, C., Yee, A., Dharamsi, A., Kluger, Y., Gerstein, M., Arrowsmith, C. H. und Edwards, A. E. (2000) Structural proteomics: prospects for high throughput sample preparation. *Prog Bioph Mol Biol* **73**: 339-45.
- Christendat, D., Yee, A., Dharamsi, A., Kluger, Y., Savchenko, A., Cort, J. R., Booth, V., Mackereth, C. D., Saridakis, V., Ekiel, I., Kozlov, G., Maxwell, K. L., Wu, N., McIntosh, L. P., Gehring, K., Kennedy, M. A., Davidson, A. R., Pai, E. F., Gerstein, M., Edwards, A. M. und Arrowsmith, C. H. (2000) Structural proteomics of an archaeon. *Nat Struct Biol* **7**: 903-9.
- Chromiak, J. A., Abadie, B. R., Braswell, R. A., Koh, Y. S. und Chilek, D. R. (2003) Resistance training exercises acutely reduce intraocular pressure in physically active men and women. *J Strength Con Res* **17**: 715-20.
- Cid-Arregui, A., Juarez, V. und zur Hausen, H. (2003) A synthetic E7 gene of human papillomavirus type 16 that yields enhanced expression of the protein in mammalian cells and is useful for DNA immunization studies. *J Virol* **77**: 4928-37.
- Cigan, A. M., Pabich, E. K. und Donahue, T. F. (1988) Mutational analysis of the HIS4 translational initiator region in *Saccharomyces cerevisiae*. *Mol Cell Biol* **8**: 2964-75.
- Coghlan, A. und Wolfe, K. H. (2000) Relationship of codon bias to mRNA concentration and protein length in *Saccharomyces cerevisiae*. *Yeast* **16**: 1131-45.
- Comeron, J. M. und Aguade, M. (1998) An evaluation of measures of synonymous codon usage bias. *J Mol Evol* **47**: 268-74.
- Comeron, J. M., Kreitman, M. und Aguade, M. (1999) Natural selection on synonymous sites is correlated with gene length and recombination in *Drosophila*. *Genetics* **151**: 239-49.
- Corbet, S., Vinner, L., Hougaard, D. M., Bryder, K., Nielsen, H. V., Nielsen, C. und Fomsgaard, A. (2000) Construction, biological activity, and immunogenicity of synthetic envelope DNA vaccines based on a primary, CCR5-tropic, early HIV type 1 isolate (BX08) with human codons. *AIDS Res Hum Retroviruses* **16**: 1997-2008.
- Couderc, R. und Baratti, J. (1980) Oxidation of Methanol by the yeast, *Pichia pastoris*. Purification and properties of the alcohol oxidase. *Agric Biol Chem* **44**: 2279-89.
- Craig, E. A., Eisenman, H. C. und Hundley, H. A. (2003) Ribosome-tethered molecular chaperones: the first line of defense against protein misfolding? *Curr Opin Microbiol* **6**: 157-62.
- Craig, K. L. und Tyers, M. (1999) The F-box: a new motif for ubiquitin dependent proteolysis in cell cycle regulation and signal transduction. *Prog Biophys Mol Biol* **72**: 299-328.
- Das, B., Guo, Z., Russo, P., Chartrand, P. und Sherman, F. (2000) The role of nuclear cap binding protein Cbc1p of yeast in mRNA termination and degradation. *Mol Cell Biol* **20**: 2827-38.
- de Smit, M. H. und van Duin, J. (1994) Control of translation by mRNA secondary structure in *Escherichia coli*. A quantitative analysis of literature data. *J Mol Biol* **244**: 144-50.
- de Smit, M. H. und van Duin, J. (1994) Translational initiation on structured messengers. Another role for the Shine-Dalgarno interaction. *J Mol Biol* **235**: 173-84.

- de Smit, M. H. und van Duin, J. (2003) Translational standby sites: how ribosomes may deal with the rapid folding kinetics of mRNA. *J Mol Biol* **331**: 737-43.
- de Smit, M. H., van Duin, J., van Knippenberg, P. H. und van Eijk, H. G. (1994) CCC.UGA: a new site of ribosomal frameshifting in *Escherichia coli*. *Gene* **143**: 43-7.
- Deml, L., Bojak, A., Steck, S., Graf, M., Wild, J., Schirmbeck, R., Wolf, H. und Wagner, R. (2001) Multiple effects of codon usage optimization on expression and immunogenicity of DNA candidate vaccines encoding the human immunodeficiency virus type 1 Gag protein. *J Virol* **75**: 10991-1001.
- Deng, T. (1997) Bacterial expression and purification of biologically active mouse c-Fos proteins by selective codon optimization. *FEBS Lett* **409**: 269-72.
- Disbrow, G. L., Sunitha, I., Baker, C. C., Hanover, J. und Schlegel, R. (2003) Codon optimization of the HPV-16 E5 gene enhances protein expression. *Virology* **311**: 105-14.
- Dobson, C. M. und Karplus, M. (1999) The fundamentals of protein folding: bringing together theory and experiment. *Curr Opin Struct Biol* **9**: 92-101.
- Dolinski, K., Balakrishnan, R., Christie, K. R., Costanzo, M. C., Dwight, S. S., Engel, S. R., Fisk, D. G., Hirschman, J. E., Hong, E. L., Issel-Tarver, L., Sethuraman, A., Theesfeld, C. L., Binkley, G., Lane, C., Schroeder, M., Dong, S., Weng, S., Andrada, R., Botstein, D. und Cherry, J. M. (2003) Saccharomyces Genome Database. **2003**.
- Doyle, S. A., Murphy, M. B., Massi, J. M. und Richardson, P. M. (2003) High-throughput proteomics: a flexible and efficient pipeline for protein production. *J Proteome Res* **1**: 531-6.
- Duret, L. (2002) Evolution of synonymous codon usage in metazoans. *Curr Opin Genet Dev* **12**: 640-9.
- Duret, L. und Mouchiroud, D. (1999) Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila*, and *Arabidopsis*. *Proc Natl Acad Sci U S A* **96**: 4482-7.
- Edwards, A. M., Arrowsmith, C. H., Christendat, D., Dharamsi, A., Friesen, J. D., Greenblatt, J. F. und Vedadi, M. (2000) Protein production: feeding the crystallographers and NMR spectroscopists. *Nat Struct Biol* **7** (Suppl): 970-2.
- Elf, J., Nilsson, D., Tenson, T. und Ehrenberg, M. (2003) Selective charging of tRNA isoacceptors explains patterns of codon usage. *Science* **300**: 1718-22.
- Ellis, S. B., Brust, P. F., Koutz, P. J., Waters, A. F., Harpold, M. M. und Gingeras, T. R. (1985) Isolation of alcohol oxidase and two other methanol regulatable genes from the yeast *Pichia pastoris*. *Mol Cell Biol* **5**: 1111-21.
- Emini, E. A., Hughes, J. V., Perlow, D. S. und Boger, J. (1985) Induction of hepatitis A virus-neutralizing antibody by a virus-specific synthetic peptide. *J Virol* **55**: 836-9.
- Emory, S. A., Bouvet, P. und Belasco, J. G. (1992) A 5'-terminal stem-loop structure can stabilize mRNA in *Escherichia coli*. *Genes Dev* **6**: 135-48.
- Faber, K. N., Harder, W., Ab, G. und Veenhuis, M. (1995) Review: methylotrophic yeasts as factories for the production of foreign proteins. *Yeast* **11**: 1331-44.
- Fahnestock, S. R. und Bedzyk, L. A. (1997) Production of synthetic spider dragline silk protein in *Pichia pastoris*. *Appl Microbiol Biotechnol* **47**: 33-9.
- Falquet, L., Pagni, M., Bucher, P., Hulo, N., Sigrist, C. J., Hofmann, K. und Bairoch, A. (2002) The PROSITE database, its status in 2002. *Nucleic Acids Res* **30**: 235-8.

- Farabaugh, P. J. (1996) Programmed translational frameshifting. *Microbiol Rev* **60**: 103-34.
- Feng, L., Chan, W. W., Roderick, S. L. und Cohen, D. E. (2000) High-level expression and mutagenesis of recombinant human phosphatidylcholine transfer protein using a synthetic gene: evidence for a C-terminal membrane binding domain. *Biochemistry* **39**: 15399-409.
- Forman, M. D., Stack, R. F., Masters, P. S., Hauer, C. R. und Baxter, S. M. (1998) High level, context dependent misincorporation of lysine for arginine in *Saccharomyces cerevisiae* a1 homeodomain expressed in *Escherichia coli*. *Protein Sci* **7**: 500-3.
- Garber, K. (2001) Biotech industry faces new bottleneck. *Nat Biotechnol* **19**: 184-5.
- Gavin, A. C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J. M., Michon, A. M., Cruciat, C. M., Remor, M., Hofert, C., Schelder, M., Brajenovic, M., Ruffner, H., Merino, A., Klein, K., Hudak, M., Dickson, D., Rudi, T., Gnau, V., Bauch, A., Bastuck, S., Huhse, B., Leutwein, C., Heurtier, M. A., Copley, R. R., Edelmann, A., Querfurth, E., Rybin, V., Drewes, G., Raida, M., Bouwmeester, T., Bork, P., Seraphin, B., Kuster, B., Neubauer, G. und Superti-Furga, G. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**: 141-7.
- Gentil Perret, A., Duthel, R., Fotso, M. J., Brunon, J. und Mosnier, J. F. (2002) Stromolysin-3 is expressed by aggressive Meningiomas. *Cancer* **94**: 765-72.
- Gilbert, M. und Albala, J. S. (2002) Accelerating code to function: sizing up the protein production line. *Curr Opin Chem Biol* **6**: 102-5.
- Gilon, T., Chomsky, O. und Kulka, R. G. (2000) Degradation signals recognized by the Ubc6p-Ubc7p ubiquitin-conjugating enzyme pair. *Mol Cell Biol* **20**: 7214-9.
- Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., Johnston, M., Louis, E. J., Mewes, H. W., Murakami, Y., Philippsen, P., Tettelin, H. und Oliver, S. G. (1996) Life with 6000 genes. *Science* **274**: 546, 63-7.
- Goh, C. S., Lan, N., Echols, N., Douglas, S. M., Milburn, D., Bertone, P., Xiao, R., Ma, L. C., Zheng, D., Wunderlich, Z., Acton, T., Montelione, G. T. und Gerstein, M. (2003) SPINE 2: a system for collaborative structural proteomics within a federated database framework. *Nucleic Acids Res* **31**: 2833-8.
- Gough, J. und Chothia, C. (2002) SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments. *Nucleic Acids Res* **30**: 268-72.
- Gough, J., Karplus, K., Hughey, R. und Chothia, C. (2001) Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J Mol Biol* **313**: 903-19.
- Gouka, R. J., Punt, P. J. und van den Hondel, C. A. (1997) Efficient production of secreted proteins by *Aspergillus*: progress, limitations and prospects. *Appl Microbiol Biotechnol* **47**: 1-11.
- Gouy, M. und Gautier, C. (1982) Codon usage in bacteria: correlation with gene expressivity. *Nucleic Acids Res* **10**: 7055-74.
- Graber, J. H., Cantor, C. R., Mohr, S. C. und Smith, T. F. (1999) In silico detection of control signals: mRNA 3'-end-processing sequences in diverse species. *Proc Natl Acad Sci U S A* **96**: 14055-60.
- Graber, J. H., McAllister, G. D. und Smith, T. F. (2002) Probabilistic prediction of *Saccharomyces cerevisiae* mRNA 3'-processing sites. *Nucleic Acids Res* **30**: 1851-8.

- Grantham, R., Gautier, C., Gouy, M., Mercier, R. und Pave, A. (1980) Codon catalog usage and the genome hypothesis. *Nucleic Acids Res* **8**: r49-r62.
- Grinna, L. S. und Tschopp, J. F. (1989) Size distribution and general structural features of N-linked oligosaccharides from the methylotrophic yeast, *Pichia pastoris*. *Yeast* **5**: 107-15.
- Griswold, K. E., Mahmood, N. A., Iverson, B. L. und Georgiou, G. (2003) Effects of codon usage versus putative 5'(-)-mRNA structure on the expression of *Fusarium solani* cutinase in the *Escherichia coli* cytoplasm. *Protein Expr Purif* **27**: 134-42.
- Grundy, W. N., Bailey, T. L., Elkan, C. P. und Baker, M. E. (1997) Hidden Markov model analysis of motifs in steroid dehydrogenases and their homologs. *Biochem Biophys Res Commun* **231**: 760-6.
- Guarna, M. M., Lesnicki, G. J., Tam, B. M., Robinson, J., Radziminski, C. Z., Hasenwinkle, D., Boraston, A., Jervis, E., MacGillivray, R. T. A., Turner, R. F. B. und Kilburn, D. G. (1997) On-Line Monitoring and Control of Methanol Concentration in Shake-Flask Cultures of *Pichia pastoris*. *Biotechnol Bioeng* **56**: 279-86.
- Guisez, Y., Tison, B., Vandekerckhove, J., Demolder, J., Bauw, G., Haegeman, G., Fiers, W. und Contreras, R. (1991) Production and purification of recombinant human interleukin-6 secreted by the yeast *Saccharomyces cerevisiae*. *Eur J Biochem* **198**: 217-22.
- Gurkan, C. und Ellar, D. J. (2003) Expression of the *Bacillus thuringiensis* Cyt2Aa1 toxin in *Pichia pastoris* using a synthetic gene construct. *Biotechnol Appl Biochem*.
- Gursky, Y. G. und Beabealashvili, R. (1994) The increase in gene expression induced by introduction of rare codons into the C terminus of the template. *Gene* **148**: 15-21.
- Gygi, S. P., Rochon, Y., Franza, B. R. und Aebersold, R. (1999) Correlation between protein and mRNA abundance in yeast. *Mol Cell Biol* **19**: 1720-30.
- Halpern, J. L., Habig, W. H., Neale, E. A. und Stibitz, S. (1990) Cloning and expression of functional fragment C of tetanus toxin. *Infect Immun* **58**: 1004-9.
- Hamdan, F. F., Mousa, A. und Ribeiro, P. (2002) Codon optimization improves heterologous expression of a *Schistosoma mansoni* cDNA in HEK293 cells. *Parasitol Res* **88**: 583-6.
- Hammarstroem, M., Hellgren, N., van den Berg, S., Berglund, H. und Haerd, T. (2002) Rapid screening for improved solubility of small human proteins produced as fusion proteins in *Escherichia coli*. *Prot Science* **11**: 313-21.
- Heinemann, U., Frevert, J., Hofmann, K., Illing, G., Maurer, C., Oschkinat, H. und Saenger, W. (2000) An integrated approach to structural genomics. *Prog Biophys Mol Biol* **73**: 347-62.
- Hellwig, S., Emde, F., Raven, N. P., Henke, M., van Der Logt, P. und Fischer, R. (2001) Analysis of single-chain antibody production in *Pichia pastoris* using on-line methanol control in fed-batch and mixed-feed fermentations. *Biotechnol Bioeng* **74**: 344-52.
- Herrick, D., Parker, R. und Jacobson, A. (1990) Identification and comparison of stable and unstable mRNAs in *Saccharomyces cerevisiae*. *Mol Cell Biol* **10**: 2269-84.
- Higgins, D. G. und Sharp, P. M. (1988) CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. *Gene* **73**: 237-44.
- Higgins, D. G., Thompson, J. D. und Gibson, T. J. (1996) Using CLUSTAL for multiple sequence alignments. *Methods Enzymol* **266**: 383-402.
- Hochstrasser, M. (1996) Protein degradation or regulation: Ub the judge. *Cell* **84**: 813-5.

- Holz, C., Prinz, B., Bolotina, N., Sievert, V., Büssow, K., Simon, B., Stahl, U. und Lang, C. (2003) Establishing the yeast *S. cerevisiae* as a system for expression of human proteins on a proteome-scale. *J Struct Funct Genomics* **4**: 97-108.
- Hoyt, M. A., Zhang, M. und Coffino, P. (2003) Ubiquitin-independent mechanisms of mouse ornithine decarboxylase degradation are conserved between mammalian and fungal cells. *J Biol Chem* **278**: 12135-43.
- Hughes, J. D., Estep, P. W., Tavazoie, S. und Church, G. M. (2000) Computational identification of cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *J Mol Biol* **296**: 1205-14.
- Humphreys, D. P., Carrington, B., Bowering, L. C., Ganesh, R., Sehdev, M., Smith, B. J., King, L. M., Reeks, D. G., Lawson, A. und Popplewell, A. G. (2002) A plasmid system for optimization of Fab' production in *Escherichia coli*: importance of balance of heavy chain and light chain synthesis. *Protein Expr Purif* **26**: 309-20.
- Ikemura, T. (1982) Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. Differences in synonymous codon choice patterns of yeast and *Escherichia coli* with reference to the abundance of isoaccepting transfer RNAs. *J Mol Biol* **158**: 573-97.
- Ikemura, T. (1985) Codon usage and tRNA content in unicellular and multicellular organisms. *Mol Biol Evol* **2**: 13-34.
- Imai, T. und Ohno, T. (1995) Measurement of yeast intracellular pH by image processing and the change it undergoes during growth phase. *J Biotechnol* **38**: 165-72.
- Invitrogen (1997) A Manual of Methods for Expression of Recombinant Proteins in *Pichia pastoris*. San Diego, Invitrogen Corporation.
- Janatova, I., Costaglioli, P., Wesche, J., Masson, J. M. und Meilhoc, E. (2003) Development of a reporter system for the yeast *Schwanniomyces occidentalis*: influence of DNA composition and codon usage. *Yeast* **20**: 687-701.
- Jansen, R., Bussemaker, H. J. und Gerstein, M. (2003) Revisiting the codon adaptation index from a whole-genome perspective: analyzing the relationship between gene expression and codon occurrence in yeast using a variety of models. *Nucleic Acids Res* **31**: 2242-51.
- Jansen, R. und Gerstein, M. (2000) Analysis of the yeast transcriptome with structural and functional categories: characterizing highly expressed proteins. *Nucleic Acids Res* **28**: 1481-8.
- Jones, D. T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol* **292**: 195-202.
- Ju, L. W., Xing, L. H., Hong, P. W. und Jin, W. J. (1998) GeneDn: for high-level expression design of heterologous genes in a prokaryotic system. *Bioinformatics* **14**: 884-5.
- Kane, J. F., Violand, B. N., Curran, D. F., Staten, N. R., Duffin, K. L. und Bogosian, G. (1992) Novel in-frame two codon translational hop during synthesis of bovine placental lactogen in a recombinant strain of *Escherichia coli*. *Nucleic Acids Res* **20**: 6707-12.
- Katz, L. und Burge, C. B. (2003) Widespread selection for local RNA secondary structure in coding regions of bacterial genes. *Genome Res* **13**: 2042-51.
- Kellis, M., Patterson, N., Endrizzi, M., Birren, B. und Lander, E. S. (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423**: 241-54.
- Kim, C. H., Oh, Y. und Lee, T. H. (1997) Codon optimization for high-level expression of human erythropoietin (EPO) in mammalian cells. *Gene* **199**: 293-301.

Kofman, A., Graf, M., Bojak, A., Deml, L., Bieler, K., Kharazova, A., Wolf, H. und Wagner, R. (2003) HIV-1 gag expression is quantitatively dependent on the ratio of native and optimized codons. *Tsitologiya* **45**: 86-93.

Kruskal, W. H. und Wallis, W. A. (1952) Use of ranks in one-criterion variance analysis. *J Amer Statist Assoc* **47**: 583-621.

Kyte, J. und Doolittle, R. F. (1982) A simple method for displaying the hydropathic character of a protein. *J Mol Biol* **157**: 105-32.

La, D., Silver, M., Edgar, R. C. und Livesay, D. R. (2003) Using motif-based methods in multiple genome analyses: a case study comparing orthologous mesophilic and thermophilic proteins. *Biochemistry* **42**: 8988-98.

Lacy-Hulbert, A., Thomas, R., Li, X. P., Lilley, C. E., Coffin, R. S. und Roes, J. (2001) Interruption of coding sequences by heterologous introns can enhance the functional expression of recombinant genes. *Gene Ther* **8**: 649-53.

Laemmli, U. K. (1970) Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature* **227**: 680-5.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczy, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chisoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M. (2001) Initial sequencing and analysis of the human genome. *Nature* **409**: 860-921.

Laroche, Y., Storme, V., De Meutter, J., Messens, J. und Lauwereys, M. (1994) High-level secretion and very efficient isotopic labeling of tick anticoagulant peptide (TAP) expressed in the methylotrophic yeast, *Pichia pastoris*. *Biotechnology (N Y)* **12**: 1119-24.

Lennon, G., Auffray, C., Polymeropoulos, M. und Soares, M. B. (1996) The I.M.A.G.E. Consortium: an integrated molecular analysis of genomes and their expression. *Genomics* **33**: 151-2.

Levy, S., Hannenhalli, S. und Workman, C. (2001) Enrichment of regulatory signals in conserved non-coding genomic sequence. *Bioinformatics* **17**: 871-7.

Lewin, B. (1990) *Genes*. Oxford, Oxford University Press.

Li, H., Ma, Y., Su, T., Che, Y., Dai, C. und Sun, M. (2003) Expression, purification, and characterization of recombinant human neuriturin secreted from the yeast *Pichia pastoris*. *Protein Expr Purif* **30**: 11-7.

Li, Z., Xiong, F., Lin, Q., d'Anjou, M., Daugulis, A. J., Yang, D. S. C. und Hew, C. L. (2001) Low-temperature increases the yield of biologically active herring antifreeze protein in *Pichia pastoris*. *Prot Express Purif* **21**: 438-45.

- Looman, A. C., Bodlaender, J., Comstock, L. J., Eaton, D., Jhurani, P., de Boer, H. A. und van Knippenberg, P. H. (1987) Influence of the codon following the AUG initiation codon on the expression of a modified lacZ gene in *Escherichia coli*. *Embo J* **6**: 2489-92.
- Looman, A. C., Laude, M. und Stahl, U. (1991) Influence of the codon following the initiation codon on the expression of the lacZ gene in *Saccharomyces cerevisiae*. *Yeast* **7**: 157-65.
- Lu, K. V., Rohde, M. F., Thomason, A. R., Kenney, W. C. und Lu, H. S. (1995) Mistranslation of a TGA termination codon as tryptophan in recombinant platelet-derived growth factor expressed in *Escherichia coli*. *Biochem J* **309** (Pt 2): 411-7.
- Ma, H. H., Yang, L., Yang, X. Y., Xu, Z. P. und Li, B. L. (2003) Bacterial expression, purification, and in vitro N-myristoylation of fusion hepatitis B virus preS1 with the native-type N-terminus. *Protein Expr Purif* **27**: 49-54.
- Maeda, Y., Kuroki, R., Suzuki, H. und Reiländer, H. (2000) High-level secretion of biologically active recombinant human macrophage inflammatory protein-1 α by the methylotrophic yeast *Pichia pastoris*. *Protein Express Purif* **18**: 56-63.
- Manns, A., Konig, H., Baier, M., Kurth, R. und Grosse, F. (1991) Fidelity of reverse transcriptase of the simian immunodeficiency virus from African green monkey. *Nucleic Acids Res* **19**: 533-7.
- Marin, A., Gallardo, M., Kato, Y., Shirahige, K., Gutierrez, G., Ohta, K. und Aguilera, A. (2003) Relationship between G+C content, ORF-length and mRNA concentration in *Saccharomyces cerevisiae*. *Yeast* **20**: 703-11.
- Martin, S. L., Vrhovski, B. und Weiss, A. S. (1995) Total synthesis and expression in *Escherichia coli* of a gene encoding human tropoelastin. *Gene* **154**: 159-66.
- Massaer, M., Mazzu, P., Haumont, M., Magi, M., Daminet, V., Bollen, A. und Jacquet, A. (2001) High-level expression in mammalian cells of recombinant house dust mite allergen ProDer p 1 with optimized codon usage. *Int Arch Allergy Immunol* **125**: 32-43.
- McCarthy, J. E. (1998) Posttranscriptional control of gene expression in yeast. *Microbiol Mol Biol Rev* **62**: 1492-553.
- McCue, L. A., Thompson, W., Carmack, C. S. und Lawrence, C. E. (2002) Factors influencing the identification of transcription factor binding sites by cross-species comparison. *Genome Res* **12**: 1523-32.
- McGrew, J. T., Leiske, D., Dell, B., Klinke, R., Krasts, D., Wee, S. F., Abbott, N., Armitage, R. und Harrington, K. (1997) Expression of trimeric CD40 ligand in *Pichia pastoris*: use of a rapid method to detect high-level expressing transformants. *Gene* **187**: 193-200.
- McGuffin, L. J., Bryson, K. und Jones, D. T. (2000) The PSIPRED protein structure prediction server. *Bioinformatics* **16**: 404-5.
- McGuffin, L. J. und Jones, D. T. (2003) Benchmarking secondary structure prediction for fold recognition. *Proteins* **52**: 166-75.
- McNulty, D. E., Claffee, B. A., Huddleston, M. J. und Kane, J. F. (2003) Mistranslational errors associated with the rare arginine codon CGG in *Escherichia coli*. *Protein Expr Purif* **27**: 365-74.
- Meetei, A. R. und Rao, M. R. (1998) Hyperexpression of rat spermatidal protein TP2 in *Escherichia coli* by codon optimization and engineering the vector-encoded 5' UTR. *Protein Expr Purif* **13**: 184-90.

- Melvin, B. K. und Shanks, J. V. (1996) Influence of aeration on cytoplasmic pH of yeast in an NMR airlift bioreactor. *Biotechnol Prog* **12**: 257-65.
- Mikaelian, I., Krieg, M., Gait, M. J. und Karn, J. (1996) Interactions of INS (CRS) elements and the splicing machinery regulate the production of Rev-responsive mRNAs. *J Mol Biol* **257**: 246-64.
- Milek, R. L., Stunnenberg, H. G. und Konings, R. N. (2000) Assembly and expression of a synthetic gene encoding the antigen Pfs48/45 of the human malaria parasite *Plasmodium falciparum* in yeast. *Vaccine* **18**: 1402-11.
- Mitchell, P. und Tollervey, D. (2000) mRNA stability in eukaryotes. *Curr Opin Genet Dev* **10**: 193-8.
- Mochizuki, S., Hamato, N., Hirose, M., Miyano, K., Ohtani, W., Kameyama, S., Kuwae, S., Tokuyama, T. und Ohi, H. (2001) Expression and characterisation of recombinant human antithrombin III in *Pichia pastoris*. *Protein Express Purif* **23**: 55-65.
- Moreno, J. I., David, N. R., Miernyk, J. A. und Randall, D. D. (2000) *Pisum sativum* mitochondrial pyruvate dehydrogenase can be assembled as a functional $\alpha 2\beta 2$ heterotetramer in the cytoplasm of *Pichia pastoris*. *Prot Expr Purif* **19**: 276-83.
- Moriyama, E. N. und Powell, J. R. (1998) Gene length and codon usage bias in *Drosophila melanogaster*, *Saccharomyces cerevisiae* and *Escherichia coli*. *Nucleic Acids Res* **26**: 3188-93.
- Mount, D. W. (2001) Bioinformatics. Cold Spring Harbor, Cold Spring Harbor Laboratory Press.
- Mwangi, S. M., Stabel, J., Lee, E. K., Kehrli, M. E. und Taylor, M. J. (2000) Expression and characterization of a recombinant soluble form of bovine tumor necrosis factor receptor type I. *Vet Immuno Immunopath* **77**: 233-41.
- Newton-Vinson, P., Hubalek, F. und Edmondson, D. E. (2000) High-level expression of human liver monoamine oxidase in *Pichia pastoris*. *Prot Expr Purif* **20**: 334-45.
- Niepel, M., Ling, J. und Gallie, D. R. (1999) Secondary structure in the 5'-leader or 3'-untranslated region reduces protein yield but does not affect the functional interaction between the 5'-cap and the poly(A) tail. *FEBS Lett* **462**: 79-84.
- Nilsson, L. O. und Mannervik, B. (2001) Improved heterologous expression of human glutathione transferase A4-4 by random silent mutagenesis of codons in the 5' region. *Biochim Biophys Acta* **1528**: 101-6.
- O'Connor, M. (1998) tRNA imbalance promotes -1 frameshifting via near-cognate decoding. *J Mol Biol* **279**: 727-36.
- Okazaki, Y., Furuno, M., Kasukawa, T., Adachi, J., Bono, H., Kondo, S., Nikaido, I., Osato, N., Saito, R., Suzuki, H., Yamanaka, I., Kiyosawa, H., Yagi, K., Tomaru, Y., Hasegawa, Y., Nogami, A., Schonbach, C., Gojobori, T., Baldarelli, R., Hill, D. P., Bult, C., Hume, D. A., Quackenbush, J., Schriml, L. M., Kanapin, A., Matsuda, H., Batalov, S., Beisel, K. W., Blake, J. A., Bradt, D., Brusic, V., Chothia, C., Corbani, L. E., Cousins, S., Dalla, E., Dragani, T. A., Fletcher, C. F., Forrest, A., Frazer, K. S., Gaasterland, T., Gariboldi, M., Gissi, C., Godzik, A., Gough, J., Grimmond, S., Gustincich, S., Hirokawa, N., Jackson, I. J., Jarvis, E. D., Kanai, A., Kawaji, H., Kawasaki, Y., Kedzierski, R. M., King, B. L., Konagaya, A., Kurochkin, I. V., Lee, Y., Lenhard, B., Lyons, P. A., Maglott, D. R., Maltais, L., Marchionni, L., McKenzie, L., Miki, H., Nagashima, T., Numata, K., Okido, T., Pavan, W. J., Perte, G., Pesole, G., Petrovsky, N., Pillai, R., Pontius, J. U., Qi, D., Ramachandran, S., Ravasi, T., Reed, J. C., Reed, D. J., Reid, J., Ring, B. Z., Ringwald, M., Sandelin, A., Schneider, C., Semple, C. A., Setou, M., Shimada, K., Sultana, R., Takenaka, Y., Taylor, M. S., Teasdale, R.

- D., Tomita, M., Verardo, R., Wagner, L., Wahlestedt, C., Wang, Y., Watanabe, Y., Wells, C., Wilming, L. G., Wynshaw-Boris, A., Yanagisawa, M. (2002) Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* **420**: 563-73.
- Oliveira, C. C., van den Heuvel, J. J. und McCarthy, J. E. (1993) Inhibition of translational initiation in *Saccharomyces cerevisiae* by secondary structure: the roles of the stability and position of stem-loops in the mRNA leader. *Mol Microbiol* **9**: 521-32.
- Orengo, C. A., Bray, J. E., Hubbard, T., LoConte, L. und Sillitoe, I. (1999) Analysis and assessment of ab initio three-dimensional prediction, secondary structure, and contacts prediction. *Proteins Suppl*: 149-70.
- Outchkourov, N. S., Stiekema, W. J. und Jongsma, M. A. (2002) Optimization of the expression of equistatin in *Pichia pastoris*. *Protein Expr Purif* **24**: 18-24.
- Pan, W., Ravot, E., Tolle, R., Frank, R., Mosbach, R., Turbachova, I. und Bujard, H. (1999) Vaccine candidate MSP-1 from *Plasmodium falciparum*: a redesigned 4917 bp polynucleotide enables synthesis and isolation of full-length protein from *Escherichia coli* and mammalian cells. *Nucleic Acids Res* **27**: 1094-103.
- Pande, S., Vimaladithan, A., Zhao, H. und Farabaugh, P. J. (1995) Pulling the ribosome out of frame by +1 at a programmed frameshift site by cognate binding of aminoacyl-tRNA. *Mol Cell Biol* **15**: 298-304.
- Peeters, K., De Wilde, C., De Jaeger, G., Angenon, G. und Depicker, A. (2001) Production of antibodies and antibody fragments in plants. *Vaccine* **19**: 2756-61.
- Pikaart, M. J. und Felsenfeld, G. (1996) Expression and codon usage optimization of the erythroid-specific transcription factor cGATA-1 in baculoviral and bacterial systems. *Protein Expr Purif* **8**: 469-75.
- Prapunwattana, P., Sirawaraporn, W., Yuthavong, Y. und Santi, D. V. (1996) Chemical synthesis of the *Plasmodium falciparum* dihydrofolate reductase-thymidylate synthase gene. *Mol Biochem Parasitol* **83**: 93-106.
- Preston, B. D., Poiesz, B. J. und Loeb, L. A. (1988) Fidelity of HIV-1 reverse transcriptase. *Science* **242**: 1168-71.
- Prinz, B., Schultchen, J., Rydzewski, R., Holz, C., Boettner, M., Stahl, U. und Lang, C. (eingereicht) Establishing a versatile fermentation and purification procedure for human proteins expressed in the yeasts *Saccharomyces cerevisiae* and *Pichia pastoris* for structural genomics. *J Struct Funct Genomics*.
- Proudfoot, N. J. und Brownlee, G. G. (1976) 3' non-coding region sequences in eukaryotic messenger RNA. *Nature* **263**: 211-4.
- Punt, P. J., van Biezen, N., Conesa, A., Albers, A., Mangnus, J. und van den Hondel, C. (2002) Filamentous fungi as cell factories for heterologous protein production. *Trends Biotechnol* **20**: 200-6.
- Rechsteiner, M. und Rogers, S. W. (1996) PEST sequences and regulation by proteolysis. *Trends Biochem Sci* **21**: 267-71.
- Reddy, T. R., Kraus, G., Suhasini, M., Leavitt, M. C. und Wong-Staal, F. (1995) Identification and mapping of inhibitory sequences in the human immunodeficiency virus type 2 vif gene. *J Virol* **69**: 5167-70.

- Ribas, A. V., Ho, P. L., Tanizaki, M. M., Raw, I. und Nascimento, A. L. (2000) High-level expression of tetanus toxin fragment C-thioredoxin fusion protein in *Escherichia coli*. *Biotechnol Appl Biochem* **31** (Pt 2): 91-4.
- Ribeiro, J. M. und Sillero, A. (1991) A program to calculate the isoelectric point of macromolecules. *Comput Biol Med* **21**: 131-41.
- Rice, P., Longden, I. und Bleasby, A. (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* **16**: 276-7.
- Richardson, J. P. (2002) Rho-dependent termination and ATPases in transcript termination. *Biochim Biophys Acta* **1577**: 251-60.
- Roberts, J. D., Bebenek, K. und Kunkel, T. A. (1988) The accuracy of reverse transcriptase from HIV-1. *Science* **242**: 1171-3.
- Roberts, J. D., Preston, B. D., Johnston, L. A., Soni, A., Loeb, L. A. und Kunkel, T. A. (1989) Fidelity of two retroviral reverse transcriptases during DNA-dependant DNA synthesis in vitro. *Mol Cell Biol* **9**: 469-76.
- Rogers, S., Wells, R. und Rechsteiner, M. (1986) Amino acid sequences common to rapidly degraded proteins: the PEST hypothesis. *Science* **234**: 364-8.
- Romanos, M. A., Makoff, A. J., Fairweather, N. F., Beesley, K. M., Slater, D. E., Rayment, F. B., Payne, M. M. und Clare, J. J. (1991) Expression of tetanus toxin fragment C in yeast: gene synthesis is required to eliminate fortuitous polyadenylation sites in AT-rich DNA. *Nucleic Acids Res* **19**: 1461-7.
- Sachs, L. (2002) Angewandte Statistik. Berlin, Springer Verlag.
- Sagliocco, F. A., Vega Laso, M. R., Zhu, D., Tuite, M. F., McCarthy, J. E. und Brown, A. J. (1993) The influence of 5'-secondary structures upon ribosome binding to mRNA during translation in yeast. *J Biol Chem* **268**: 26522-30.
- Saris, N., Holkeri, H., Craven, R. A., Stirling, C. J. und Makarow, M. (1997) The Hsp70 homologue Lhs1p is involved in a novel function of the yeast endoplasmic reticulum, refolding and stabilization of heat-denatured protein aggregates. *J Cell Biol* **137**: 813-24.
- Satchidanandam, V. und Shivashankar, Y. (1997) Availability of a second upstream AUG can completely overcome inhibition of protein synthesis initiation engendered by mRNA secondary structure encompassing the start codon. *Gene* **196**: 231-7.
- Sawasaki, T., Ogasawara, T., Morishita, R. und Endo, Y. (2002) A cell-free protein synthesis system for high-throughput proteomics. *Proc Natl Acad Sci U S A* **99**: 14652-7.
- Scheich, C., Sievert, V. und Buessow, K. (2003) An automated method for high-throughput protein purification applied to a comparison of His-tag and GST-tag affinity chromatography. *BMC Biotech* **3**: 12-20.
- Schultz, J., Copley, R. R., Doerks, T., Ponting, C. P. und Bork, P. (2000) SMART: a web-based tool for the study of genetically mobile domains. *Nucleic Acids Res* **28**: 231-4.
- Schwartz, R., Ting, C. S. und King, J. (2001) Whole proteome pI values correlate with subcellular localizations of proteins for organisms within the three domains of life. *Genome Res* **11**: 703-9.
- Schwartz, S., Campbell, M., Nasioulas, G., Harrison, J., Felber, B. K. und Pavlakis, G. N. (1992) Mutational inactivation of an inhibitory sequence in human immunodeficiency virus type 1 results in Rev-independent gag expression. *J Virol* **66**: 7176-82.

- Schwartz, S., Felber, B. K. und Pavlakis, G. N. (1992) Distinct RNA sequences in the gag region of human immunodeficiency virus type 1 decrease RNA stability and inhibit expression in the absence of Rev protein. *J Virol* **66**: 150-9.
- Scorer, C. A., Buckholz, R. G., Clare, J. J. und Romanos, M. A. (1993) The intracellular production and secretion of HIV-1 envelope protein in the methylotrophic yeast *Pichia pastoris*. *Gene* **136**: 111-9.
- Seetharam, R., Heeren, R. A., Wong, E. Y., Braford, S. R., Klein, B. K., Aykent, S., Kotts, C. E., Mathis, K. J., Bishop, B. F., Jennings, M. J. und et al. (1988) Mistranslation in IGF-1 during over-expression of the protein in *Escherichia coli* using a synthetic gene containing low frequency codons. *Biochem Biophys Res Commun* **155**: 518-23.
- Seffens, W. und Digby, D. (1999) mRNAs have greater negative folding free energies than shuffled or codon choice randomized sequences. *Nucleic Acids Res* **27**: 1578-84.
- Senejani, A. G., Hilario, E. und Gogarten, J. P. (2001) The intein of the *Thermoplasma* A-ATPase A subunit: structure, evolution and expression in *E. coli*. *BMC Biochem* **2**: 13.
- Sharp, P. M. und Cowe, E. (1991) Synonymous codon usage in *Saccharomyces cerevisiae*. *Yeast* **7**: 657-78.
- Sharp, P. M. und Li, W. H. (1987) The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res* **15**: 1281-95.
- Sharp, P. M., Tuohy, T. M. und Mosurski, K. R. (1986) Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Res* **14**: 5125-43.
- Sillero, A. und Ribeiro, J. M. (1989) Isoelectric points of proteins: theoretical determination. *Anal Biochem* **179**: 319-25.
- Sinclair, G. und Choy, F. Y. (2002) Synonymous codon usage bias and the expression of human glucocerebrosidase in the methylotrophic yeast, *Pichia pastoris*. *Protein Expr Purif* **26**: 96-105.
- Slavik, J. und Kotyk, A. (1984) Intracellular pH distribution and transmembrane pH profile of yeast cells. *Biochim Biophys Acta* **766**: 679-84.
- Slimko, E. M. und Lester, H. A. (2003) Codon optimization of *Caenorhabditis elegans* GluCl ion channel genes for mammalian cells dramatically improves expression levels. *J Neurosci Methods* **124**: 75-81.
- Spanjaard, R. A. und van Duin, J. (1988) Translation of the sequence AGG-AGG yields 50% ribosomal frameshift. *Proc Natl Acad Sci U S A* **85**: 7967-71.
- States, D. J. und Botstein, D. (1991) Molecular sequence accuracy and the analysis of protein coding regions. *Proc Natl Acad Sci USA* **88**: 5518-22.
- Stenstrom, C. M., Jin, H., Major, L. L., Tate, W. P. und Isaksson, L. A. (2001) Codon bias at the 3'-side of the initiation codon is correlated with translation initiation efficiency in *Escherichia coli*. *Gene* **263**: 273-84.
- Stryer, L. (1991) Biochemie. Heidelberg, Spektrum Akademischer Verlag.
- Talarico, L. A., Ingram, L. O. und Maupin-Furlow, J. A. (2001) Production of the Gram-positive *Sarcina ventriculi* pyruvate decarboxylase in *Escherichia coli*. *Microbiology* **147**: 2425-35.
- Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J. und Church, G. M. (1999) Systematic determination of genetic network architecture. *Nat Genet* **22**: 213-5.

- Te'o, V. S., Cziferszky, A. E., Bergquist, P. L. und Nevalainen, K. M. (2000) Codon optimization of xylanase gene *xynB* from the thermophilic bacterium *Dictyoglomus thermophilum* for expression in the filamentous fungus *Trichoderma reesei*. *FEMS Microbiol Lett* **190**: 13-9.
- Terwilliger, T. C., Park, M. S., Waldo, G. S., Berendzen, J., Hung, L. W., Kim, C. Y., Smith, C. V., Sacchettini, J. C., Bellinzoni, M., Bossi, R., De Rossi, E., Mattevi, A., Milano, A., Riccardi, G., Rizzi, M., Roberts, M. M., Coker, A. R., Fossati, G., Mascagni, P., Coates, A. R., Wood, S. P., Goulding, C. W., Apostol, M. I., Anderson, D. H., Gill, H. S., Eisenberg, D. S., Taneja, B., Mande, S., Pohl, E., Lamzin, V., Tucker, P., Wilmanns, M., Colovos, C., Meyer-Klaucke, W., Munro, A. W., McLean, K. J., Marshall, K. R., Leys, D., Yang, J. K., Yoon, H. J., Lee, B. I., Lee, M. G., Kwak, J. E., Han, B. W., Lee, J. Y., Baek, S. H., Suh, S. W., Komen, M. M., Arcus, V. L., Baker, E. N., Lott, J. S., Jacobs, W., Jr., Alber, T. und Rupp, B. (2003) The TB structural genomics consortium: a resource for *Mycobacterium tuberculosis* biology. *Tuberculosis (Edinb)* **83**: 223-49.
- Tregoning, J. S., Nixon, P., Kuroda, H., Svab, Z., Clare, S., Bowe, F., Fairweather, N., Ytterberg, J., van Wijk, K. J., Dougan, G. und Maliga, P. (2003) Expression of tetanus toxin Fragment C in tobacco chloroplasts. *Nucleic Acids Res* **31**: 1174-9.
- Tschopp, J. F., Brust, P. F., Cregg, J. M., Stillman, C. A. und Gingeras, T. R. (1987) Expression of the *lacZ* gene from two methanol-regulated promoters in *Pichia pastoris*. *Nucleic Acids Res* **15**: 3859-76.
- Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S. und Rothberg, J. M. (2000) A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* **403**: 623-7.
- van den Heuvel, J. J., Planta, R. J. und Raue, H. A. (1990) Effect of leader primary structure on the translational efficiency of phosphoglycerate kinase mRNA in yeast. *Yeast* **6**: 473-82.
- Vega Laso, M. R., Zhu, D., Sglicioco, F., Brown, A. J., Tuite, M. F. und McCarthy, J. E. (1993) Inhibition of translational initiation in the yeast *Saccharomyces cerevisiae* as a function of the stability and position of hairpin structures in the mRNA leader. *J Biol Chem* **268**: 6453-62.
- Vincentelli, R., Bignon, C., Gruez, A., Canaan, S., Sulzenbacher, G., Tegoni, M., Campanacci, V. und Cambillau, C. (2003) Medium-scale structural genomics: strategies for protein expression and crystallization. *Acc Chem Res* **36**: 165-72.
- Wang, H., O'Mahony, D. J., McConnell, D. J. und Qi, S. Z. (1993) Optimization of the synthesis of porcine somatotropin in *Escherichia coli*. *Appl Microbiol Biotechnol* **39**: 324-8.
- Weik, R., Francky, A., Striedner, G., Raspor, P., Bayer, K. und Mattanovich, D. (1998) Recombinant expression of alliin lyase from garlic (*Allium sativum*) in bacteria and yeasts. *Planta Med* **64**: 387-8.
- Weissman, A. M. (2001) Themes and variations on ubiquitylation. *Nat Rev Mol Cell Biol* **2**: 169-78.
- Welihinda, A. A., Tirasophon, W. und Kaufman, R. J. (1999) The cellular response to protein misfolding in the endoplasmic reticulum. *Gene Expr* **7**: 293-300.
- Winzeler, E. A., Shoemaker, D. D., Astromoff, A., Liang, H., Anderson, K., Andre, B., Bangham, R., Benito, R., Boeke, J. D., Bussey, H., Chu, A. M., Connolly, C., Davis, K., Dietrich, F., Dow, S. W., El Bakkoury, M., Foury, F., Friend, S. H., Gentalen, E., Giaever, G.,

- Hegemann, J. H., Jones, T., Laub, M., Liao, H., Davis, R. W. und et al. (1999) Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science* **285**: 901-6.
- Withers-Martinez, C., Carpenter, E. P., Hackett, F., Ely, B., Sajid, M., Grainger, M. und Blackman, M. J. (1999) PCR-based gene synthesis as an efficient approach for expression of the A+T-rich malaria genome. *Protein Eng* **12**: 1113-20.
- Woo, J. H., Liu, Y. Y., Mathias, A., Stavrou, S., Wang, Z., Thompson, J. und Neville, D. M., Jr. (2002) Gene optimization is necessary to express a bivalent anti-human anti-T cell immunotoxin in *Pichia pastoris*. *Protein Expr Purif* **25**: 270-82.
- Wright, F. (1990) The 'effective number of codons' used in a gene. *Gene* **87**: 23-9.
- Xu, Z. L., Mizuguchi, H., Ishii-Watabe, A., Uchida, E., Mayumi, T. und Hayakawa, T. (2001) Optimization of transcriptional regulatory elements for constructing plasmid vectors. *Gene* **272**: 149-56.
- Yelin, R. und Schuldiner, S. (2001) Vesicular monoamine transporters heterologously expressed in the yeast *Saccharomyces cerevisiae* display high-affinity tetrabenazine binding. *Biochim Biophys Acta* **1510**: 426-41.
- Yokoyama, S. (2003) Protein expression systems for structural genomics and proteomics. *Curr Opin Chem Biol* **7**: 39-43.
- Zdanovsky, A. G. und Zdanovskaia, M. V. (2000) Simple and efficient method for heterologous expression of clostridial proteins. *Appl Environ Microbiol* **66**: 3166-73.
- Zhang, H., Howard, E. M. und Roepe, P. D. (2002) Analysis of the antimalarial drug resistance protein Pfert expressed in yeast. *J Biol Chem* **277**: 49767-75.
- Zhang, S. P., Zubay, G. und Goldman, E. (1991) Low-usage codons in *Escherichia coli*, yeast, fruit fly and primates. *Gene* **105**: 61-72.
- Zhang, W., Bevins, M. A., Plantz, B. A., Smith, L. A. und Meagher, M. M. (2000) Modeling *Pichia pastoris* growth on methanol and optimizing the production of a recombinant protein, the heavy-chain fragment C of botulinum neurotoxin, serotype A. *Biotechnol Bioeng* **70**: 1-8.
- Zhang, W., Hywood Potter, K. J., Plantz, B. A., Schlegel, V. L., Smith, L. A. und Meagher, M. M. (2003) *Pichia pastoris* fermentation with mixed-feeds of glycerol and methanol: growth kinetics and production improvement. *J Ind Microbiol Biotechnol* **30**: 210-5.
- Zhao, J., Hyman, L. und Moore, C. (1999) Formation of mRNA 3' ends in eukaryotes: mechanism, regulation, and interrelationships with other steps in mRNA synthesis. *Microbiol Mol Biol Rev* **63**: 405-45.
- Zhou, R., Kroczyńska, B. und Miernyk, J. A. (2000) Expression of the *Arabidopsis thaliana* AtJ2 cochaperone protein in *Pichia pastoris*. *Prot Expr Purif* **19**.

Eigene Veröffentlichungen:

- Boettner, M., Steffens, C. Stahl, U. und Lang, C. (in Vorbereitung) Sequence based factors influencing the expression of heterologous genes in the yeast *Pichia pastoris* – a comparative view on 79 human genes.
- Prinz, B., Schultchen, J., Rydzewski, R., Holz, C., Boettner, M., Stahl, U. und Lang, C. (eingereicht) Establishing a versatile fermentation and purification procedure for human

proteins expressed in the yeasts *Saccharomyces cerevisiae* and *Pichia pastoris* for structural genomics. *J Struct Funct Genomics*.

Boettner, M. und Lang, C. (im Druck) High throughput expression in microplate formate in *Pichia pastoris*. In: Recombinant Gene Expression Protocols, 2nd Edition.

Boettner, M., Prinz, B., Holz, C., Stahl, U. und Lang, C. (2002) High-throughput screening for expression of heterologous proteins in the yeast *Pichia pastoris*. *J Biotechnol* **99**: 51-62.

C. Lang, T. Polakowski, B. Prinz, M. Boettner, U. Stahl (1999) Heterologous membrane proteins in *Sacharomyces cerevisiae*: intracellular targeting of bacterioopsin. In: Endocytobiology VII (E. Wagner et al., Eds.), University of Geneva, 397 – 405

Posterpräsentationen:

High-throughput screening and expression of heterologous proteins in the yeast *Pichia pastoris*.

M. Boettner, C. Steffens und C. Lang; 2nd International Conference on Recombinant Protein Production 2002, Cernobbio, Italien

High-throughput screening for expression of heterologous proteins in the yeast *Pichia pastoris*.

M. Boettner, C. Steffens, Krause, T. und C. Lang; International Conference on Structural Genomics 2002, Berlin

High-throughput expression and purification of heterologous proteins in *Pichia pastoris*.

M. Boettner, B. Prinz, C. Holz, C. Lang; XXth International Conference on Yeast Genetics and Molecular Biology 2001, Prag, Tschechische Republik

Optimizing the growth and feeding strategy for the expression of heterologous proteins in *Pichia pastoris*.

M. Boettner, C. Holz, C. Lang; Molecular Biology of Yeasts and Related Organisms 2000, Ober-Ramstadt

High-throughput Expression of Human Genome Sequences in Yeast.

C. Holz, M. Boettner, N. Bolotina, O. Hesse, C. Lang; Biotechnology 2000, Berlin

Internationale Vorträge:

High-throughput expression and purification of heterologous proteins in *Pichia pastoris*.

M. Boettner, B. Prinz, C. Holz, C. Lang; XXth International Conference on Yeast Genetics and Molecular Biology 2001, Prag, Tschechische Republik

8 Anhang

Tabelle 1: Übersicht über die analysierten cDNAs
Integrale Membranproteine (siehe 4.3.2.2) sind fett gedruckt.

interne ID	GenBank ProteinID	Länge der cDNA [Bp]	Protein [kDa]	interne ID	GenBank ProteinID	Länge der cDNA [Bp]	Protein [kDa]
Nuller							
500000279	BAA03400	426	15,8	500000373	AAB38529	627	23,2
500000043	AAD34115	528	19,7	500000375	AAC51284	312	12,1
500000282	AAD34095	684	26	500000383	AAC27445	342	12,8
500000049	AAA93231	477	18,1	500000386	AAD03265	567	21
500000114	AAC34987	654	24,7	500000390	AAD16169	627	22,9
500000155	AAA74903	201	7,9	500000392	CAA36235	327	11,9
500000286	CAA22906	123	4,6	500000394	AAB64192	504	19,7
500000289	AAB81453	489	18,3	500000397	CAA30792	714	25,6
500000296	CAB56506	297	10,3	500000398	AAD11629	741	28,2
500000298	AAA35648	1425	51,7	500000406	CAA51827	357	12,9
500000304	AAF15100	249	9,3	500000408	AAA59982	684	25,4
500000313	AAD27769	657	24,8	500000410	AAA35709	711	26
500000317	AAB23825	801	29,6	500000412	AAA20587	1431	52,8
500000321	AAD20972	282	10,8	500000417	AAA18209	495	17,7
500000322	AAC39912	981	38	500000422	AAD34133	459	17,6
500000325	CAA58535	1194	45,6	500000425	BAA05118	240	9,3
500000327	AAB88175	693	24,4	500000430	AAA70088	270	10,1
500000328	AAD44492	426	16,7	500000438	BAA33391	474	18,7
500000333	CAA27385	909	35,1	500000440	AAC35550	519	19,7
500000335	AAB53091	927	33,3	500000443	AAA61187	702	25,5
500000339	BAA11485	978	37,8	500000448	AAA63269	276	10
500000340	AAD08720	1029	38,9	500000450	AAB24206	327	11,7
500000342	AAD21526	348	13,5	500000452	AAD49967	1116	43
500000350	AAB25225	132	5	500000463	AAA36597	732	27,4
500000354	CAA45277	636	24,4	500000467	AAC25187	654	23,4
500000356	AAF03512	510	19,9	500000469	AAD34064	1053	38,6
500000358	CAA65339	252	9,2				
500000361	CAB52345	876	33,9				
500000365	AAC18356	300	10,7				
500000371	AAA87064	759	28,3				

interne ID	GenBank ProteinID	Länge der cDNA [Bp]	Protein [kDa]	interne ID	GenBank ProteinID	Länge der cDNA [Bp]	Protein [kDa]
Einser							
500000112	AAA60286	354	13,3	500000385	AAB81205	456	17,8
500000156	AAD09623	1011	38,2	500000400	CAB56611	1086	40,1
500000294	AAF14857	375	14,2	500000433	AAF13149	774	28,7
500000308	AAD44363	462	16,6	500000446	BAA09317	339	12,5
500000310	AAD51801	153	5,8				
500000318	AAF03537	465	17,9	Dreier			
500000348	AAA87395	816	30,6	500000048	AAD25021	438	15,9
500000379	AAD44489	606	22,9	500000344	AAF00499	663	25,4
500000389	BAA08392	357	13,4	500000362	AAC83329	477	17,1
500000405	AAB00114	999	37,5	gezielt umklonierte cDNAs:			
500000427	AAC39715	243	8,9	500001103	AAC27445	345	12,8
500000419	BAA13402	990	37,5	500001047	AAD27777	402	15,24
500000436	AAF14868	714	26,9	500001443	AAB96936	537	20,3
500000461	AAD02685	876	33,4				
500000471	AAD20048	888	32,9				
500000473	AAF14877	669	24,3				
Zweier							
500000278	AAC28637	939	34				
500000281	AAD34115	528	19,7				
500000283	AAC52115	696	27,2				
500000113	CAA34890	681	26,5				
500000290	AAD27741	819	29,3				
500000293	AAC41945	609	23,2				
500000300	CAB56175	468	17,7				
500000302	AAD54939	849	30,8				
500000307	BAA12872	600	22,9				
500000330	AAC62536	609	23,2				
500000353	CAA34200	516	19,6				

Tabelle 2: Die gefundenen Motive nach MEME

	ID	Start	'E-value'	Sequenz des Motives
Motiv 1	408-.pro	62	1.69e-49	ILIGAGALMMLVGLGCCGAVQESQCMGLFFGFLVIFAIAAAIWGY
	410-.pro	58	5.08e-49	ILIAVGAIMILGFLGCCGAIKESRCMLLFFIGLLUILLQVATGILGA
	316-.pro	58	5.01e-43	VFIGVGAVTMLMGFLGCIGAVNEVRCILGLYFAFLUUAQVTAGALFY
	396-.pro	56	7.84e-40	VIAVGVFLFLVAFVGCCGACKENYCLMITFAFLSLJMLVEVAAAAGY
	370-.pro	63	3.93e-38	ILVAGTVMVTVGLGCCATFKERRNLLRLYFILLIIFLEIAGILAY
Motiv 2	298-.pro	152	6.35e-64	PYKEMINDAAMFYTNRLKEYKDVKHMDWWKAYLSWTELQAYIKEFH
	412-.pro	154	1.03e-62	PYKEMINDAATFYTNRLKDYKHSDLRMDWWKSYLNWSELQAYIKEHH
Motiv 3	316-.pro	214	3.48e-38	EGCMEKVQAWLQENLGIILGVGVGVAMVELGMVLSCLCRHMH
	370-.pro	206	5.83e-38	GGCTIKLETFQEHLMGAVGIGIACVQVFGMIFTCCLYRSLK
	408-.pro	179	8.81e-37	KSCFDAIKEVFDNKFHIGAVGIGIAWMIFGMIFSMILCCAIR
	396-.pro	189	2.11e-35	EGCVEKIGGWLRKNMLWAAAALGIAFVELGIVFACCLVKSIR
	410-.pro	192	7.58e-34	ETCISFKDFLAKNLIMGISFGLAMELGLVFSMMLYCQIG
Motiv 4	408-.pro	6	1.14e-33	GTKCIKYLFGFNFIWLAGIAVLAIGLW
	370-.pro	12	3.65e-28	GTVCLKYLFTYNCCFWLAGLAWMAVGW
	410-.pro	4	1.50e-26	VSACIKYSMTFNFLFMLCGIILALAW
	316-.pro	5	8.54e-21	CIKVTKYFLFLNLIFFILGAVILGFGWW
	332-.pro	29	6.80e-18	GGSMFKILLFYMFYGC LAGIFGTIQVM
	396-.pro	6	2.32e-17	GMKCVKFLYMLLAFCACAVGLIIVGVG
Motiv 5	412-.pro	371	6.11e-62	NSIIDNCKKLGLVFDNVGIVEVINSQDIQIQVMGRVPTISNKTEGCH
	298-.pro	368	2.10e-61	NSTVDNCKKLGLVFDNVGIVEVINSKDVQVMGKVPITISNKIDGCH
Motiv 6	340-.pro	128	4.79e-35	CYLPHMMMQLNLLLEEGGLVQVESNLQ
	340-.pro	92	1.13e-32	CYLPHMMMQLNLLLEEDGLVQLETNLQ
	440-.pro	2	5.33e-21	CYQGYGYPLMLFLEEGGWTVCKINTQ
Motiv 7	412-.pro	96	1.11e-60	QQPHENDVAALLKPISEKIQEQTFRERNRGSNMFNHL SAVSESIPALGW
	298-.pro	94	3.08e-57	QQPAENKLSDLLAPISEKIQEVTFREKNRGSKLFNHL SAVSESIPALGW
Motiv 8	412-.pro	329	2.73e-50	GKKWRVEYQEDRNDLVISETLKQVAYIFKCEKSTIQKKG
	298-.pro	326	1.26e-47	GKKWRVENQENVSNLMIEDTELKQVAYIYKCVNTILQIKKG
Motiv 9	298-.pro	253	6.61e-45	SRSALFAQINQGESTHALKHVSDDMKTHKNPALKAQ
	412-.pro	259	1.01e-43	SRSALFAQLNQGEATKGLRHVTDDQKTYKNPSLRAQ
Motiv 10	412-.pro	422	1.64e-30	YLS DALDC EIVSAKSSEMNIIP
	298-.pro	419	6.98e-29	YLSKNSLDC EIVSAKSSEMNIIP
Motiv 11	298-.pro	38	1.31e-45	AGAAPYVQAFDSL AGPVAEYLKSK EGGDVQKHAE MVH
	412-.pro	40	6.97e-45	AGVAPSVEAFDKLMDSMVAEFLKNSRLAGDVETHAE MVH

TAGATCTAATCAAGAGGATGTCAGAATGCCATTGCGCTGAGAGATGCAGGC**TTCATTTT**GATCT**TTTTTATTT**GTAACT**TATATAGTAT**AGGAT**TTTTTT**TGCA

Abbildung 28: Die 3'-UTR der *AOX1*-Expressionskassette

Das Stop-Codon ist kursiv dargestellt, putative AT-reiche Regionen sind fett und unterstrichen.

PERL Skript zum Kruskal-Wallis Test:

```

use strict;
use vars qw ($eingabedatei $ausgabedatei $Ri0
$Ri1 $Ri2 $Ri3 $Ri0sqr $Ri1sqr $Ri2sqr $Ri3sqr
$ni0 $ni1 $ni2 $ni3 $nges $H $lfdNr $Parameter
$Klasse $vorgaenger $rangzahlzaehler
$rangzahlnerner $ranggroesse $rangzahl $element
$liste $listenplatz $table @lysAnzahl @liste @table
@table2 @line @lysAnzahl @feld0 @feld1 @feld2
@feld3);
#fragt Input Datei ab
print "welche Input Datei?\n";
$_=1;
$eingabedatei = <STDIN>;
chomp $eingabedatei;
until ((-e $eingabedatei) and (-s $eingabedatei))
{
    print "Die Datei existiert nicht\n";
    print "neuer Versuch:\n";
    $eingabedatei = <STDIN>;
    chomp $eingabedatei;
}
#fragt output ab
print "Ausgabedatei: \n";
$ausgabedatei = <STDIN>;
chomp $ausgabedatei;

open (INPUT, "$eingabedatei");

open (OUTPUT, ">ablage.txt");

while (<INPUT>) {
    my $zeile = $_;
    #kreiert eine tabgetrennte Liste der Variablen
    $zeile =~
        /([\d\\.\\]*)t([\d\\.\\]*)t([\d\\.\\]*)/;
    my ($lfdNr, $Parameter, $Klasse) =
        ($1,$2,$3);
    #kreiert eine Liste der Parameter
    @liste = ($lfdNr, $Parameter, $Klasse);
    #stellt die Referenz der Liste an das Ende von
    @table => Liste von Referenzen
    push @table, [@liste];
}

foreach $element (@table) {
    #dereferenziert einzelne Zeilen aus @table
    @line = @{$element};
    #Erzeugung von @lysAnzahl bestehend aus
    den Anzahlen der Lysine
    #2. Element von @line
    my $lysAnzahl = $line[1];
    push @lysAnzahl, $lysAnzahl;
}
#hochzählen der ranggroessen
$listenplatz = 1;
$ranggroesse = 1;
$vorgaenger = 0;
#Liste wird von Eintrag 2 bis 79 durchgenudelt
#ranggroesse wird bestimmt und an letzten Eintrag des

```

```

jeweiligen Ranges
#in @table angehängt (!referenziert)
for ($listenplatz=1; $listenplatz < 79;
$listenplatz++) {
    if ($lysAnzahl[$listenplatz] ==
        $lysAnzahl[$vorgaenger]) {
        $ranggroesse++;
        #Sonderfall, falls for Schleife in der letzten
        Zeile ist
        if ($listenplatz == 79) {
            #ranggroesse wird an @table gepusht; jeweils
            hinter dem
            #letzten Eintrag des entsprechenden Ranges
            (vorgaenger)
            my @line = @{$table[$listenplatz]};
            push @line, $ranggroesse;
            $table[$listenplatz] = [@line];
        }
    }
    else {
        #ranggroesse wird an @table gepusht; jeweils
        hinter dem
        #letzten Eintrag des entsprechenden Ranges
        (vorgaenger)
        my @line = @{$table[$vorgaenger]};
        push @line, $ranggroesse;
        $table[$vorgaenger] = [@line];
        #zurücksetzen von ranggroesse für
        nächsten Rang
        $ranggroesse = 1;
        #Sonderfall, falls for Schleife in der
        letzten Zeile ist
        if ($listenplatz == 79) {
            #ranggroesse wird an @table gepusht;
            jeweils hinter dem
            #letzten Eintrag des entsprechenden
            Ranges (vorgaenger)
            my @line = @{$table[$listenplatz]};
            push @line, $ranggroesse;
            $table[$listenplatz] = [@line];
        }
    }
    $vorgaenger++;
}
#bestimmung von rangzahl
$rangzahlzaehler = 0;
foreach $element (@table) {
    #dereferenziert einzelne Zeilen aus @table
    @line = @{$element};
    $lfdNr = $line[0];
    if ($line[3]) {
        $rangzahlzaehler =
            ($rangzahlzaehler + $lfdNr);
        $rangzahlnerner = $line[3];
        $rangzahl = ($rangzahlzaehler /
            $rangzahlnerner);
        $rangzahlzaehler = 0;
        #print OUTPUT "$rangzahl\n";
        push @line, $rangzahl;
    }
    else {
        $rangzahlzaehler =
            $rangzahlzaehler + $lfdNr;
    }
}
print OUTPUT (join "\t", @line);

```

```

    print OUTPUT "\n";
}
# close input and output file
close INPUT;
close OUTPUT;
#neuer array mit zeilen der ranglistendatei
open INPUT, "ablage.txt";
open OUTPUT, ">$ausgabedatei";
while (<INPUT>) {
    my $zeile = $_;
    #kreiert eine tabgetrennte Liste der Variablen
    if ($zeile =~
/([d\.\,]*)\t([d\.\,]*)\t([d\.\,]*)\t([d\.\,]*)\t([d\.\,]*)\t([d\.\,]*)/) {
        ($lfdNr, $Parameter, $Klasse, $ranggroesse,
        $rangzahl) = ($1,$2,$3,$4,$5);
        @liste = ($lfdNr, $Parameter,
        $Klasse,$ranggroesse, $rangzahl);
    }
    elsif ($zeile
=~ /([d\.\,]*)\t([d\.\,]*)\t([d\.\,]*)/) {
        ($lfdNr, $Parameter, $Klasse) =
        ($1,$2,$3);
        @liste = ($lfdNr, $Parameter, $Klasse);
    }
    #stellt die Referenz der Liste an das Ende von @table
    => Liste von Referenzen
    push @table2, [@liste];
}
#@table wird von hinten durchlaufen, rangzahl und
ranggroesse gemerkt und - wenn nicht vorhanden -
#an die Liste darüber angefügt
for ($listenplatz=78; $listenplatz >= 0;
$listenplatz--) {
    @line = @{$table2[$listenplatz]};
    if ($line[4]) {
        $ranggroesse = $line[3];
        $rangzahl = $line[4];
    }
    else {
        push @line, $ranggroesse;
        push @line, $rangzahl;
    }
    splice (@table2, $listenplatz, 1, [@line]);
}
foreach $element (@table2) {
    #dereferenziert einzelne Zeilen aus @table
    @line = @{$element};
    #erzeugt rangtabelle.txt
    print OUTPUT (join "\t", @line);
    print OUTPUT "\n";
}
print OUTPUT "\n";
#Erstellen von vier Tabellen mit den Einträgen der
jeweiligen Expressionsstärke
foreach $element (@table2) {
    @line = @{$element};
    if ($line[2]==0) {
        push @feld0, [@line];
    }
    elsif ($line[2]==1) {
        push @feld1, [@line];
    }
    elsif ($line[2]==2) {
        push @feld2, [@line];
    }
    elsif ($line[2]==3) {
        push @feld3, [@line];
    }
}

```

```

}
#ausgabe der Tabellen, die nur aus den
Klassenmitgliedern bestehen
foreach $element (@feld0) {
    #dereferenziert einzelne Zeilen aus @table
    @line = @{$element};
    #erzeugt rangtabelle.txt
    print OUTPUT (join "\t", @line);
    print OUTPUT "\n";
}
print OUTPUT "\n";

foreach $element (@feld1) {
    #dereferenziert einzelne Zeilen aus @table
    @line = @{$element};
    #erzeugt rangtabelle.txt
    print OUTPUT (join "\t", @line);
    print OUTPUT "\n";
}
print OUTPUT "\n";

foreach $element (@feld2) {
    #dereferenziert einzelne Zeilen aus @table
    @line = @{$element};
    #erzeugt rangtabelle.txt
    print OUTPUT (join "\t", @line);
    print OUTPUT "\n";
}
print OUTPUT "\n";

foreach $element (@feld3) {
    #dereferenziert einzelne Zeilen aus @table
    @line = @{$element};
    #erzeugt rangtabelle.txt
    print OUTPUT (join "\t", @line);
    print OUTPUT "\n";
}
print OUTPUT "\n";
#berechnen und Ausgabe von Ri, RiQuadrat und ni aus
den vier Rangtafeln

#Expr 0
$Ri0 = 0;
foreach $element (@feld0) {
    @line = @{$element};
    $Ri0 = ($Ri0 + $line[4]);
}
print "Ri0: $Ri0\t";
print OUTPUT "Ri0: $Ri0\t";
$Ri0sqr = $Ri0**2;
print "Ri0sqr: $Ri0sqr\t";
print OUTPUT "Ri0sqr: $Ri0sqr\t";
$ni0 = ($#feld0+1);
print "ni0: $ni0\n";
print OUTPUT "ni0: $ni0\n";
print OUTPUT "\n";

#Expr 1
$Ri1 = 0;
foreach $element (@feld1) {
    @line = @{$element};
    $Ri1 = ($Ri1 + $line[4]);
}
print "Ri1: $Ri1\t";
print OUTPUT "Ri1: $Ri1\t";
$Ri1sqr = $Ri1**2;
print "Ri1sqr: $Ri1sqr\t";
print OUTPUT "Ri1sqr: $Ri1sqr\t";

```

```

$ni1 = ($#feld1+1);
print "ni1: $ni1\n";
print OUTPUT "ni1: $ni1\n";
print OUTPUT "\n";

```

#Expr 2

```

$Ri2= 0;
foreach $element (@feld2) {
    @line = @{$element};
    $Ri2 = ($Ri2 + $line[4]);
}
print "Ri2: $Ri2\t";
print OUTPUT "Ri2: $Ri2\t";
$Ri2sqr = $Ri2**2;
print "Ri2sqr: $Ri2sqr\t";
print OUTPUT "Ri2sqr: $Ri2sqr\t";
$ni2 = ($#feld2+1);
print "ni2: $ni2\n";
print OUTPUT "ni2: $ni2\n";
print OUTPUT "\n";

```

#Expr 3

```

$Ri3= 0;
foreach $element (@feld3) {
    @line = @{$element};
    $Ri3 = ($Ri3 + $line[4]);
}
print "Ri3: $Ri3\t";
print OUTPUT "Ri3: $Ri3\t";
$Ri3sqr = $Ri3**2;
print "Ri3sqr: $Ri3sqr\t";
print OUTPUT "Ri3sqr: $Ri3sqr\t";

```

```

$ni3 = ($#feld3+1);
print "ni3: $ni3\n";
print OUTPUT "ni3: $ni3\n";
print OUTPUT "\n";

```

#berechnung von Summe von Risqr/ni

```

my $Ridurchn = (($Ri0sqr/$ni0) +
($Ri1sqr/$ni1) + ($Ri2sqr/$ni2) +
($Ri3sqr/$ni3));
print "Ridurchn: $Ridurchn\n";
print OUTPUT "Ridurchn: $Ridurchn\n";

```

#berechnung von nges

```

$nges=$ni0 + $ni1 + $ni2 + $ni3;
print "nges: $nges\n";
print OUTPUT "nges: $nges\n";

```

#berechnung von H

```

$H=(((12/($nges*($nges+1)))*$Ridurchn)-
(3*($nges+1)));
print "H: $H\n";
print OUTPUT "H: $H\n";

```

close input and output file

```

close INPUT;
close OUTPUT;

```